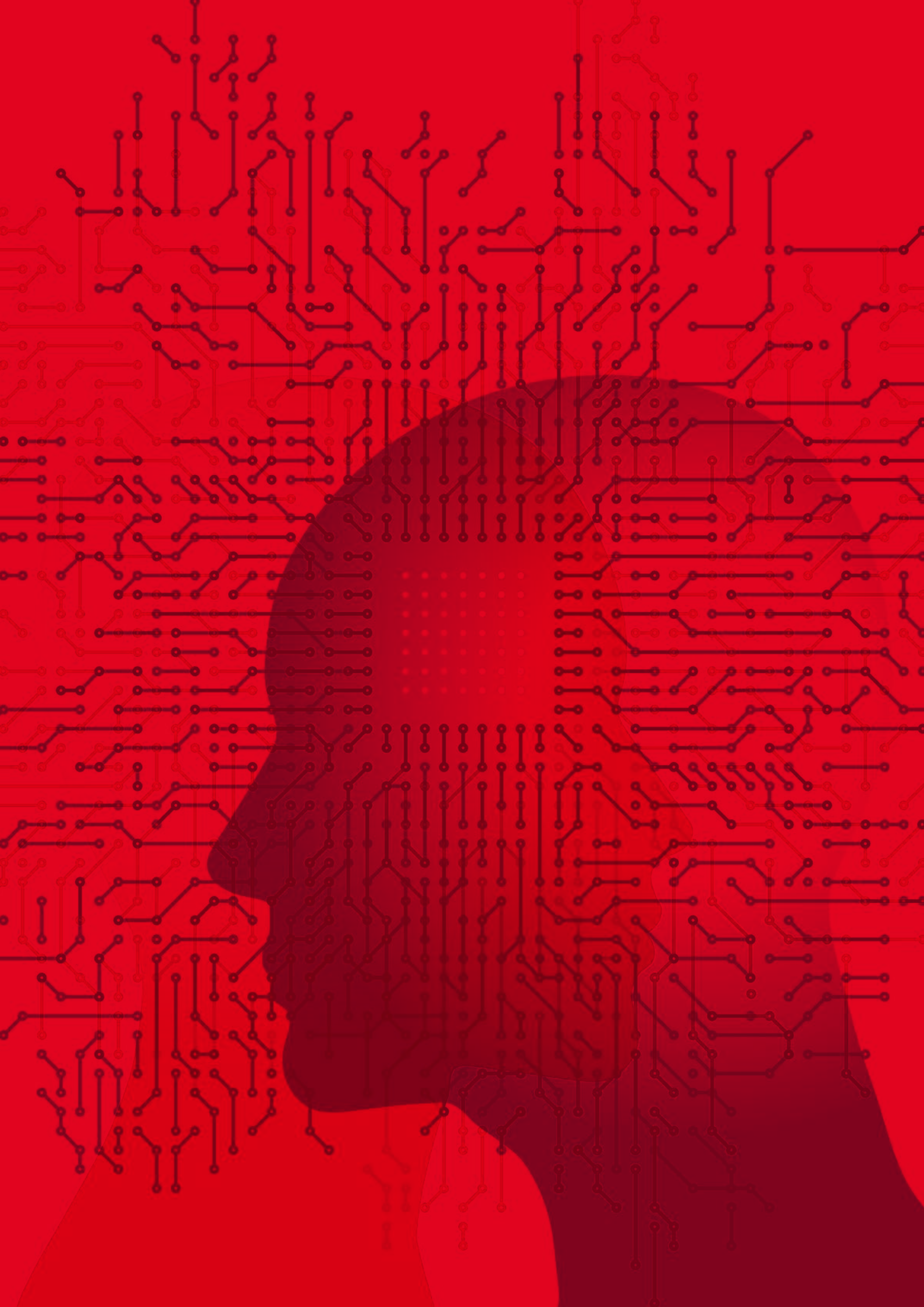
ARTIFICIAL INTELLIGENCE

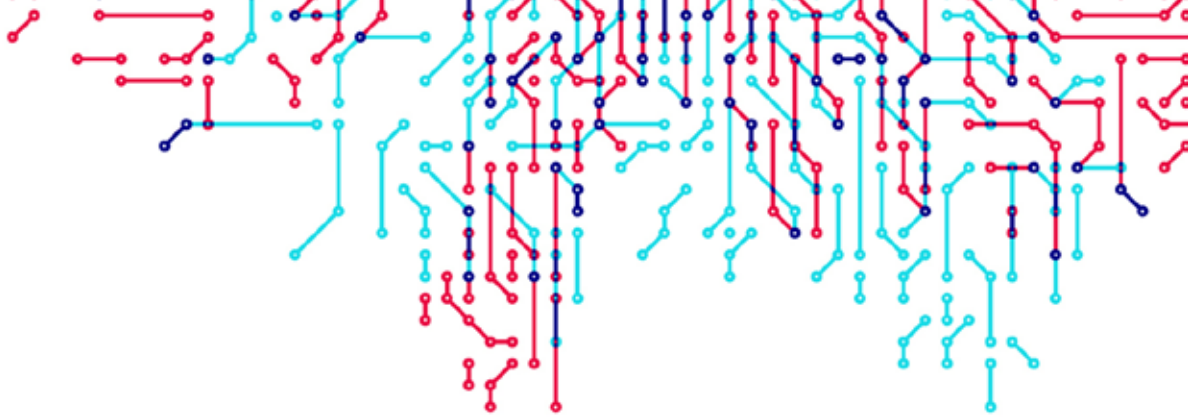
Automated Decision-making in Catalonia

apdcat
Autoritat Catalana de Protecció de Dades

Generalitat de Catalunya

Barcelona, January 2020
LD: B 2966-2020
Design and layout: Official Journal and Publications Agency





FOREWORD

Towards a smart society

By M. Àngels Barbarà,
Director of the Catalan Data Protection Authority

The last ten years have seen technological evolution so wide-ranging, so deep and so fast that we humans are unable to foresee how these technologies will be brought into our lives.[1]

The growth of innovation and technological development is founded on personal and non-personal data. The convergence of several technologies over time and their integration to use data on a mass scale and generate value greatly increases the impact on people's lives. Moreover, the exploitation of information does not drain its value in one use but rather it can be employed several times, thus building a value chain which affords enormous power to the people who control the technology and data.

I am talking about technologies such as cloud computing, which allows access from various places and huge and cheap storage capacity; the Internet of Things (IoT), which gathers information from all environments in any format and at any time; big data analytics, which makes it possible to generate new knowledge in health, transport, education, the environment, etc.; and finally artificial intelligence (AI), which refers to systems that display intelligent behaviour by analysing their environment and taking actions with some degree of autonomy to achieve specific goals.[2]

In its broadest sense, AI can be viewed as a support tool for decision-making in any field. This support for decision-making significantly affects people since data and technology profile and draw patterns which are used to take decisions and directly impact them.

1 "We have lost the ability to censure, to govern or even to negotiate with technology, which is seen as a logical consequence of our intellectual progress" (Joaquín David Rodríguez Álvarez. *La civilización ausente*. 2016, p. 11).

2 Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe. COM/2018/237 final.

Invisible changes

We are stepping into a digital setting which leaves us exposed to new forms of vulnerability through the processing of our personal data. These data are used by emerging technologies to generate value and to a great extent they change the way in which people and organisations understand the world around them.

This change is taking place almost invisibly. It is coming into our personal, professional, social and political lives as if it were something quite normal and convenient, enhancing our lives with no impact or consequences.

This is one of the biggest risks arising from emerging technologies: their silent encroachment on our rights and freedoms, transforming the society in which we live in a way which seems to be impact-free. This technology-driven society, which we call the *digital society*, will be very different from the one we have known to date.

In the digital society which is unfolding as part of the fourth industrial revolution, technologies are merging through the physical, digital and biological realms.[3] Technology is integrated into our bodies and evolves based on the new data we generate. It is designed to deepen knowledge of our drivers with the aim of influencing decision-making by and about people directly, invisibly and in real time.

Its invisibility and lack of transparency means that we are unaware of the threat it poses to the values and principles which have underpinned our society until now: dignity, freedom, free development of personality and equality. It is reformulating crucial issues in democracy and the rule of law.

New model of society

A new model of society is emerging without its members realising it. Consequently, at the Catalan Data Protection Authority and using the same approach as all European supervisory authorities and the main institutions supporting human rights, we aim to provide Catalan society with clarity and understanding about this issue. Likewise, we seek to help people safeguard their rights and freedoms more actively through knowledge and information.

On this occasion we are looking at artificial intelligence (AI) in Catalonia, and in particular automated decision-making algorithms (ADAs) because of their potential impact on privacy and all the rights and freedoms of individuals.

3 Klaus Schwab. *La cuarta revolución industrial*. Barcelona: Debate, 2016.

It goes without saying that we are doing this on the basis of data protection regulations. Since May 2018, the General Data Protection Regulation (GDPR) has been fully in force. This EU regulation is directly applicable to all European Union Member States and gives legal form to the commitment to protect the rights and freedoms of individuals as part of the shift in our model of society towards the digital society. It is also a European boost for emerging technologies as a means to lift Europe out of recession.

The GDPR lays down a new model for securing the rights and freedoms of individuals based on the accountability and commitment of organisations to the rights of individuals and on a risk-based approach in which each organisation has to evaluate how the specific data processing it performs affects its users.

This new European framework is rooted in the traditional principles of data protection (fairness, lawfulness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, security) and in requirements for the lawful basis of processing which are very similar to existing ones. However, it fundamentally changes the way in which compliance with them is addressed, precisely on the basis of the aforementioned principles of accountability and a risk-based approach.

The new model forces organisations to think about the most appropriate way to conduct their processing in order to lessen its impact on people's rights while also fully tailoring it to each organisation's specific features.

Ethical and social aspect

In recent years, the ethical and social aspect has taken on particular prominence in the debate about the evolution of the right to data protection and how this aspect should govern technological innovation and protect the values, principles and rights which are inherent in our society and identify the European tradition.

The continuous innovation and disruption environment of AI and ADAs requires organisations seeking to use these tools to go one step further in compliance. It is no longer enough to meet obligations merely mechanically and formally. These tools have to be designed, developed and used with an ethical and social dimension. Intensive use of mostly personal data makes this essential from a legal and ethical standpoint. AI cannot be an end in itself, but rather a means to achieve a transformation of society which is centred on the person. Yet accomplishing this is not easy, since AI is routinely and invisibly embedded in the technologies we use.

AI is used for many different purposes and can improve people's wellbeing. As noted by the OECD in a recent recommendation,[4] AI has the potential to contribute to global and sustainable economic activity and increase innovation and productivity.

Against this backdrop, the right to personal data protection has become a crucial factor in protecting individuals against potential abuse of information. We live in a society in which each individual's digital identity, whether or not it matches reality, is used as the basis for personalising products and services, for influencing people's decisions and for making decisions about them. Yet as the group of experts on ethics set up by the European Data Protection Supervisor points out, data protection is not a technical or legal issue but rather a profoundly human one.

Technology improves our quality of life and makes it more comfortable for us, but this cannot be in exchange for our rights being curtailed unchecked; for example, by manipulating and distorting our perception of reality or "emotional surveillance".[5] As the *Automating Society* report by Algorithm Watch notes, other questions need to be asked such as who decides when AI should be used, for what purposes and/or who develops it and how.

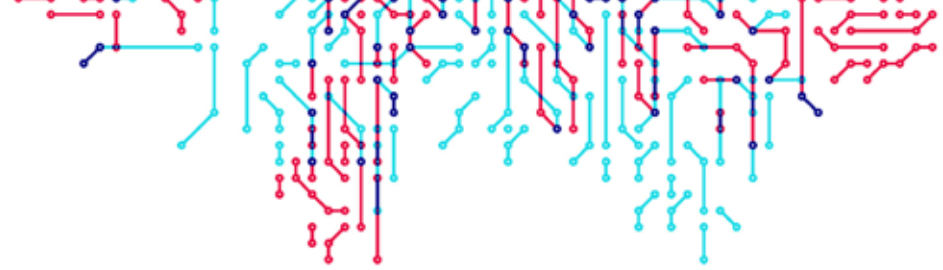
These are fundamental issues and the European Parliament underscored this when it said that AI research should invest not only in AI technology and innovation developments but also in AI-related social, ethical and accountability issues. There are factors which have to be built into the design, development and application of technology. These factors should go beyond the obligations covered by regulations and have a universal reach. In other words, they must be globally understandable and applicable.

Safe and trustworthy AI

Given these enormous capabilities of emerging technologies, we need to consider whether the pathway taken by innovation in designing and using them can be made less intrusive in people's lives.

We should drive design which does not restrict the individual's freedom to make decisions and does not allow anyone wielding power over technology to control or direct people's behaviour. We will only be able to do this if society grasps how algorithms work and can make critical judgements about the technologies used. Safe and trustworthy AI should be designed which includes the principles of transparency, explainability, security, auditability and accountability.

4 OECD, Recommendation of the Council on Artificial Intelligence (OECD/Legal/0449), adopted 21 May 2019.
5 Resolution 2018/2088(INI) of the European Parliament of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics.



In October 2018, the 40th International Conference of Data Protection and Privacy Commissioners adopted the Declaration on Ethics and Data Protection in Artificial Intelligence. In April 2019, the European Commission's independent High-Level Expert Group on Artificial Intelligence issued its ethical guidelines for trustworthy AI. In May 2019, the OECD published its recommendations on AI. Finally, at the 41st International Conference of Data Protection and Privacy Commissioners, the Council of Europe announced the start of work on the development of a legal framework for AI.

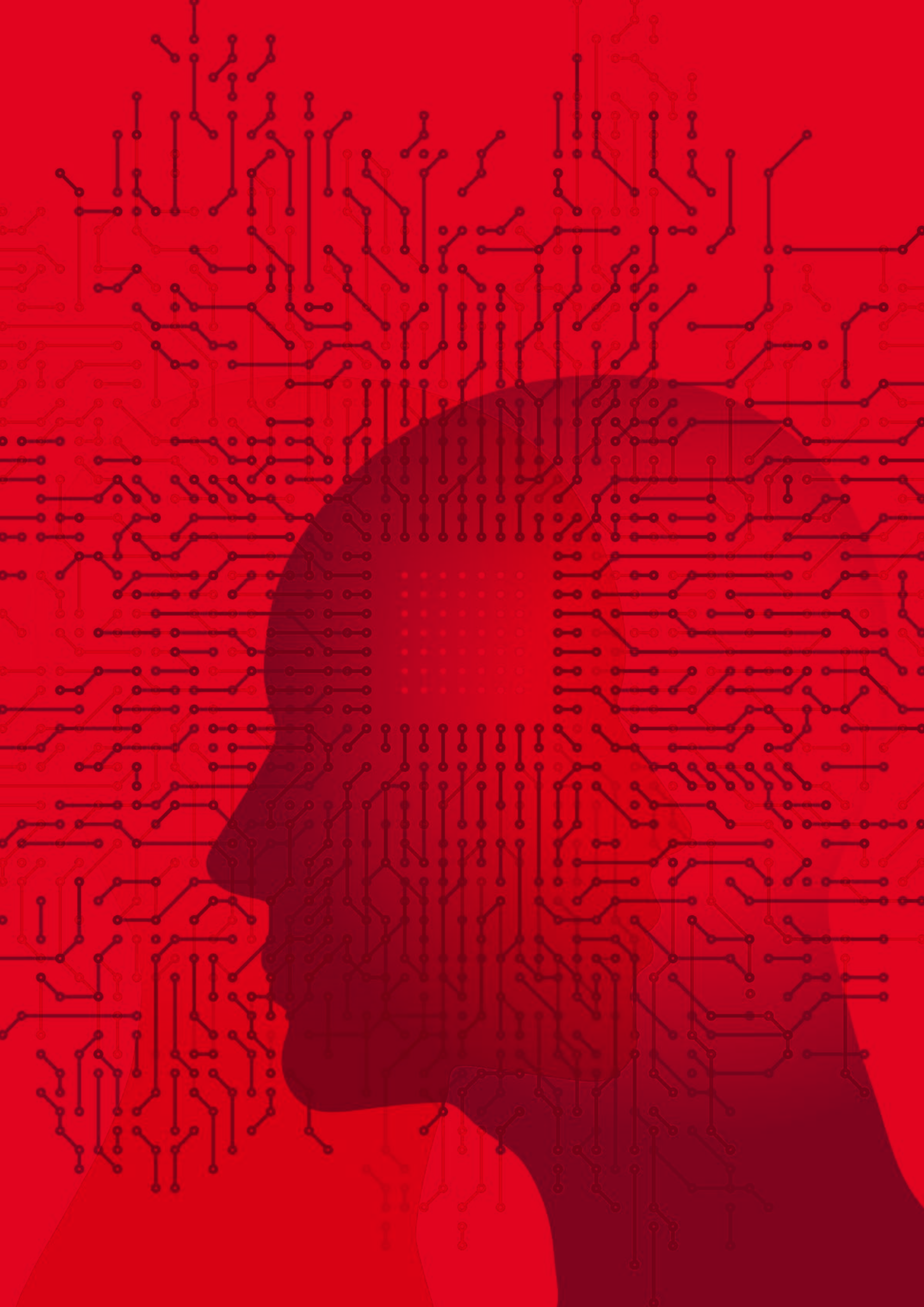
Against this backdrop and given that Catalonia is a major driving force in technological development, the Catalan Data Protection Authority has started up a specific project based on the work we have conducted in recent years in ethics and data protection which is designed to identify how AI is used in Catalonia and how the ethical aspect is built into these settings.

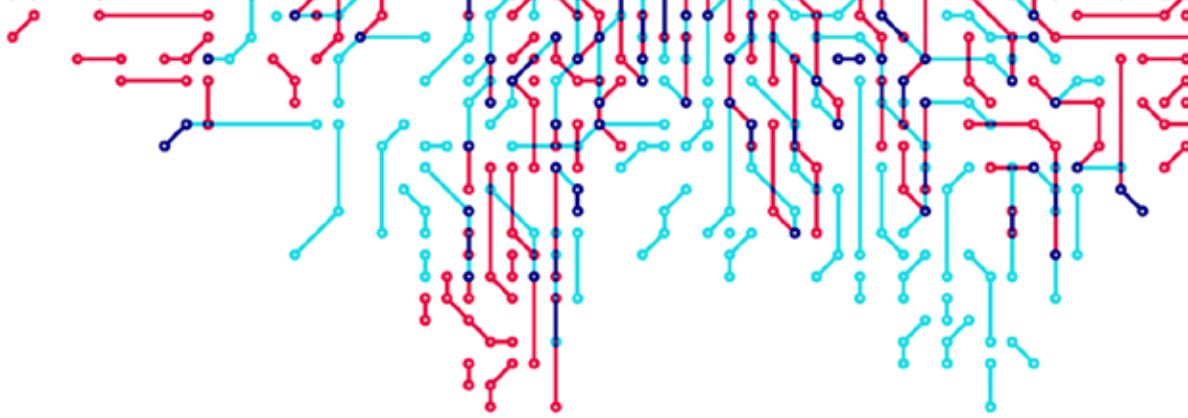
This project has focused on a specific use of AI: automated decision-making algorithms (ADAs) and how they impact the values that define the European tradition. This report includes a compilation of ADA uses by journalist Karma Peiró, a specialist in information and communication technology (ICT). I would like to point out that although the entire report is intended to be informative, this is particularly true of the description of ADA uses. To make these technologies more comprehensible, we have avoided the more technical details of their design and use.

Potential deviations and discrimination resulting from the use of emerging technologies also have to be identified so as to avoid the perpetuation of "historical" racial or sexual discrimination and the generation of new, non-perceptible discriminations rooted in the creation of groups on the basis of algorithmic patterns and correlations.

Europe's ethical approach to AI enhances people's trust and aims at building a competitive advantage for European AI companies.[6] The Catalan Data Protection Authority seeks to contribute to drawing up the principles which are to govern the design, development and use of AI. This report is only the beginning of the long road ahead of us in safeguarding the rights and freedoms of individuals in a constantly changing world.

6 Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Building Trust in Human-Centric Artificial Intelligence. COM/2019/168 final.





1.
DATA PROTECTION AND ETHICS IN AUTOMATED DECISION-MAKING

Our lives are increasingly shaped by our interaction with technology. At the heart of this ecosystem is our data. Instagram has our photos, Google our email and documents, Facebook knows our social circle and Twitter our interests. In these cases we generate the data ourselves and provide it to numerous organisations. On other occasions, the data are generated and collected without our knowledge; for example, the trail we leave with mobile devices (fitted with all kinds of sensors), Internet browsing and credit card purchases.

This huge amount of personal data available heightens the impact of technology on people's lives. In other words, it allows a very high degree of personalisation in the services offered to us. This can clearly be seen in the case of news. Previously, we could choose a newspaper or a news programme based on its editorial stance. Now there are mobile apps, news recommenders and social media (Google News, Facebook, Twitter) that tailor the information to each person's profile. The idea is to enhance users' experience by showing them only the information which is of interest to them in order to build their loyalty.

This very high degree of personalisation can have many advantages, yet it also brings with it risks in terms of rights and freedoms. The risks directly related to personal data processing are quite obvious: a security incident can jeopardise the confidentiality of our data, information can be misused, etc. Another type of not so obvious risk comes from the specific use of the data. In the case of personalised news, there is a danger that the information shown to us is too restricted, limiting our view of reality. This is what is

called the *filter bubble*. Moreover, it should also be borne in mind that the person who controls what you see can also control what you think, feel and do.

Against this background, the ability of algorithms to make automated decisions is crucial. A news portal could manually customise content for a small number of people but this task is unmanageable when users are counted in the thousands. Today, it is algorithms that analyse a user's profile and decide which content is most suitable for them.

We have given the example of a news portal, but this is only one instance of the thousands of uses that automated decision-making algorithms have in many areas. This report includes fieldwork on the use of automated decision-making algorithms which sets out many uses in healthcare, social issues, employment, banking, etc.

These uses show that automated decision-making algorithms can have a huge impact on people: mistaken diagnosis of an illness, refusal of social assistance, rejection in a staff selection process, denial of a bank loan, etc.

This report

The purpose of this report is to raise awareness and give the public the knowledge they need to use their personal data responsibly. It does this in two stages.

The first stage is research fieldwork about the use of automated decision-making algorithms in Catalonia by journalist Karma Peiró, a specialist in information and communication technology. The idea is to provide the public with information about the risks and rewards involved.

In the second stage and on the grounds that most of these applications use personal data, we look at the main conflicts in automated decision-making algorithms (and by extension artificial intelligence) along with the principles of personal data protection. Based on this discussion we then put forward a number of final recommendations.

Automated decision-making algorithms in the GDPR

Since May 2018, the General Data Protection Regulation (GDPR) has been the basic European standard governing personal data processing. The previous regulation was too formalistic and unable to address the significant challenges which have gone hand in hand with technological development. A risk-based approach and the accountability principle are two important new features of the GDPR; the aim is to adapt them to the GDPR so it can successfully meet the challenges posed by technological development. In the risk-based approach, the measures taken by the controller to safeguard the rights and freedoms of individuals are intended to be proportionate to the risks of the processing. The accountability principle means that the controller has to be able to show that their processing is in line with the GDPR. In other words, it is not enough to comply with the GDPR; you also have to be able to prove it.

The GDPR steps up the rights we have as individuals about the processing of our data; what is called informational self-determination. However, people can only exercise the rights they have over their data appropriately if they first know what these rights are and are then aware of the need to safeguard them. For example, one of the big mistakes we make is that we give our consent for the use of our data too generously or even carelessly. By always consenting to the use of our data we significantly reduce the protection the GDPR affords us.

The GDPR's wording is designed to be as general as possible. Accordingly the GDPR does not mention specific data processing and instead is based on general principles. The only reference to a specific type of processing is in Article 22 which addresses automated decision-making. The enormous importance that this type of algorithm has taken on and the major consequences it may have for individuals mean that specific safeguards have to be put in place.

Automated decision-making algorithms and artificial intelligence

Automated decision-making algorithms and artificial intelligence are closely related concepts. Obviously it is possible to design an algorithm that makes automated decisions in ways which are not very (or not at all) intelligent. For example, an algorithm for choosing the best candidate for a job might decide at random without evaluating any of the candidates. However, there would be a high likelihood that the decision made by such an algorithm would not be ideal. For automated decision-making algorithms to be really useful, they need to behave intelligently.

AI includes algorithms whose purpose is to furnish computers with intelligent behaviour. The definition of *intelligent behaviour* is not entirely clear. Nowadays AI has outstripped

human capabilities in numerous specific tasks yet it still cannot perform many of the functions of humans. In other words, it can beat the world's best chess player but it would be unable to perform the ordinary tasks of a middle-aged person such as having a meeting with their children's teacher. People move in different contexts and environments and decide on myriad issues every day because we have built up prior knowledge which machines do not yet have. This can also be explained in the following way: humans do many things with little precision; machines, very few with great precision.

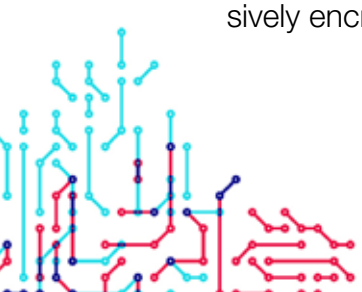
There are several levels of intelligent behaviour: from the most basic systems, which simply apply a pre-established set of rules, to complex deep learning systems, which are inspired by biological neural networks. Setting the limit beyond which an algorithm is intelligent is a contentious issue and has no special relevance for the purpose of automated decision-making algorithms, i.e. making decisions without human intervention. Accordingly, in this report we will use the terms *artificial intelligence* and *automated decision-making algorithms* interchangeably. The only requirement we set for qualifying as artificial intelligence is that the algorithm analyses the current scenario and makes its decision based on this analysis.

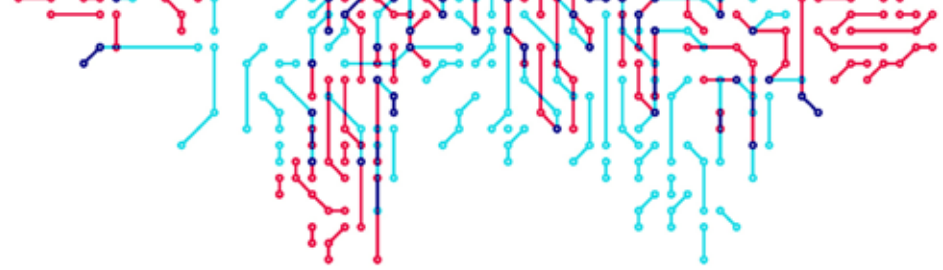
Deciding the limits and setting what is ethical

Given the major consequences that automated decision-making algorithms and artificial intelligence have for people, the limit of what is reasonable has to be set. Obviously, the law is a basic instrument, particularly the GDPR as the main European regulation on data protection. However, it is not the only instrument. Ethics also plays an essential role, for example when drawing up the law and guiding its application.

The decision as to what is and is not ethical is not an issue which can be resolved by the Catalan Data Protection Authority (APDCAT) or a group of academics. Ethics is a social contract and it is society which has to determine what is ethical. For this conversation to take place successfully, society has to be provided with the fundamental knowledge it needs. This report may be useful in this respect inasmuch as it sets out the uses of automated decision-making algorithms in Catalonia and analyses their implications.

Nowadays society views AI and ADAs in two ways. On the one hand, we are used to the improvements brought about by technological progress and we take them on board quickly; hence it could be said that not using the capabilities of artificial intelligence in detecting diseases, for example, would be unethical. On the other hand, we should not be naive and say that anything goes; being too permissive can lead to uses that excessively encroach on people's rights and freedoms.





The enormous potential of artificial intelligence is a crucial factor in setting the limits. It is seen as disruptive technology which can change power relations between states and this means no one wants to be left behind. There is already talk of the new space race between superpowers, this time anchored in the development of AI. The United States, asserting its position as the leader in technology development, still holds a dominant position. China has come in more recently with a multi-billion dollar investment plan with which in a decade it wants first to match the United States in knowledge and then outdo it. Japan is a leader in robotics. Some European countries are looking for their own piece of the artificial intelligence turf albeit on more modest budgets. It is in the development of ethical AI (also from the standpoint of data protection) that Europe is ahead of the other players.

Data protection and AI

The large amount of available data has been one of the cornerstones for the development of AI which we have seen over the last decade. Hence the impact of data protection on AI is a controversial issue.

Some sectors criticise the safeguards provided by the GDPR because they believe that they hamper the development of AI in Europe. These stakeholders advocate looser regulation which makes it easier to reuse data.

There is also the opposite view whereby the restrictions imposed by the GDPR are an incentive for the development of artificial intelligence which uses available data more reasonably and effectively, for example by encouraging the development of more sophisticated algorithms which learn from limited datasets just as people do. Chess (even though it does not concern personal data) is a very straightforward case: the most sophisticated algorithms have trained by playing billions of games while human grandmasters have played at most a few thousand.

The Catalan Data Protection Authority's standpoint is of course that the use of artificial intelligence must be compatible with the GDPR.

The society we want to build

Without playing down the criticisms of the GDPR, looking at the past can help us think about what kind of society we want to build. The first mention of the right to privacy in the context of protecting people from the intrusions of the technological advances of the time dates from 1890:[7]

7 Warren and Brandeis. "The Right to Privacy". *Harvard Law Review* (1890).

> "Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that 'what is whispered in the closet shall be proclaimed from the house-tops'."

To all intents and purposes we are faced with a similar case: deciding whether to let technology use personal data indiscriminately or whether to set limits. New technology is undoubtedly much more intrusive than a photograph insofar as it can generate a much more extensive profile of the person. If taking a photograph of a person in their private life can breach the right to privacy, should we not also limit the use of the data we generate? Article 12 in the Universal Declaration of Human Rights recognises the right to privacy.[8]

In today's world there is a wide range of views on the relationship between artificial intelligence and respect for the rights and freedoms of individuals. In particular, in terms of personal data protection we have these three scenarios:

Europe

It has the most advanced personal data protection legislation. With a few exceptions (such as legal obligations, public interest, legitimate interest), European legislation promotes informational self-determination; that is, the right of each individual to decide what their data should be used for.

As mentioned above, European data protection law has some conflicts with artificial intelligence. (*See chapter 3, "Data Protection and AI".*)

China

Personal data protection is an underdeveloped issue in China. There is a personal data protection standard but it is not compulsory.

China uses artificial intelligence in extremely intrusive ways. For example, the social credit system which is currently being introduced uses mass electronic surveillance to monitor people's behaviour, assign a score to each individual, and reward or penalise them. Some of the potential consequences of this system for Chinese citizens include travel bans, exclusion from certain schools, repression of religious minorities and public humiliation.

8 Universal Declaration of Human Rights (https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf).

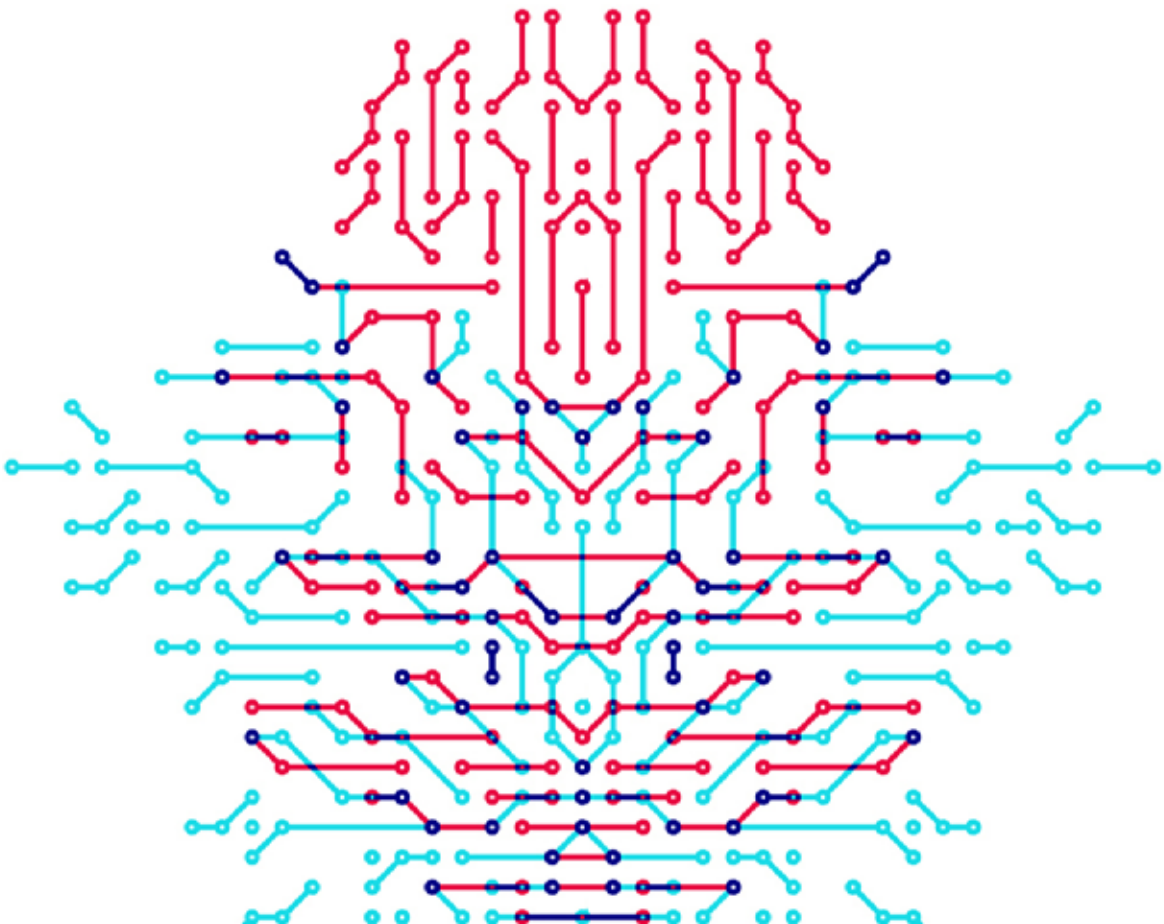
United States

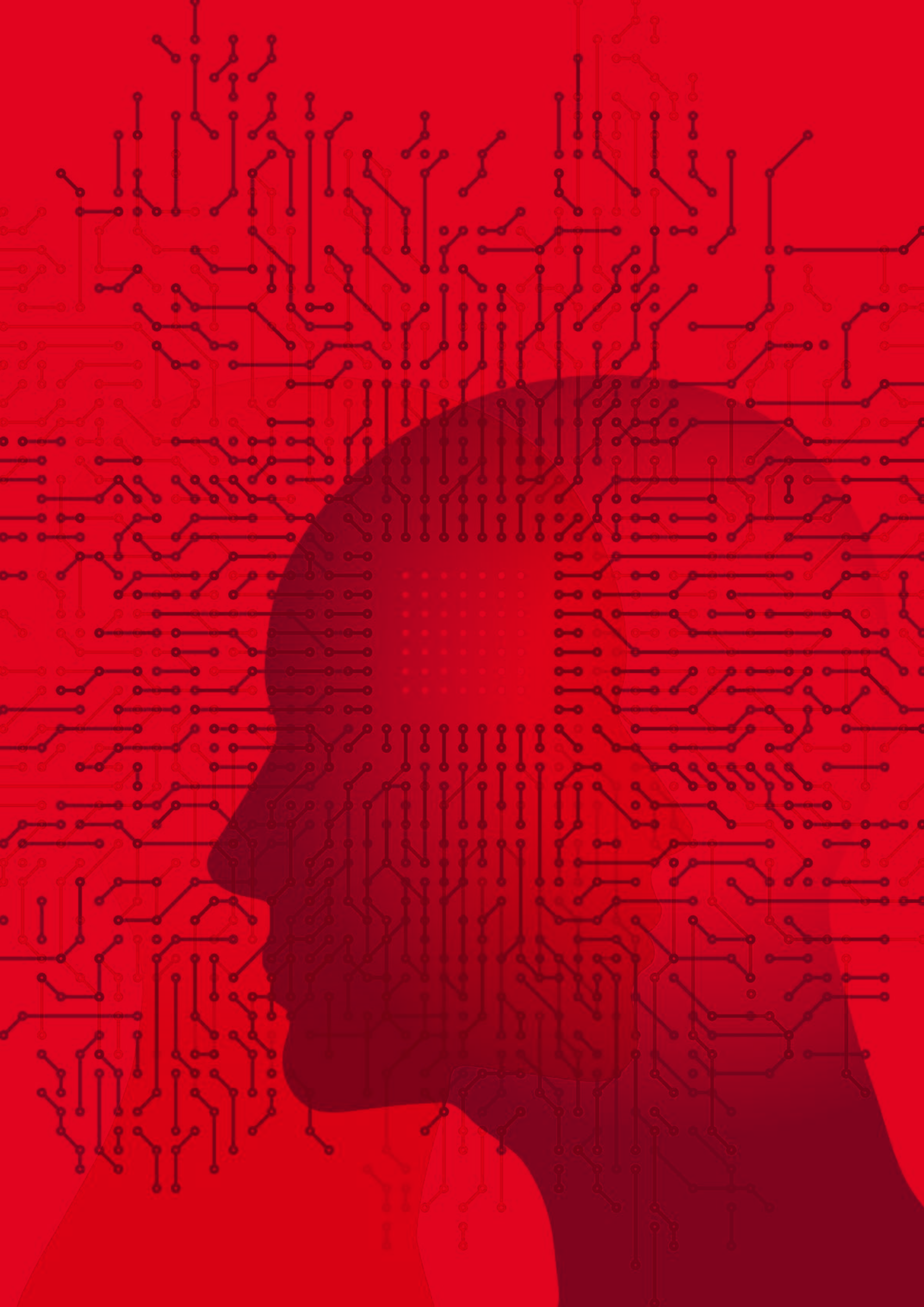
There is no one data protection law in the United States. Its legal framework is made up of hundreds of state and federal laws which deal with specific aspects.

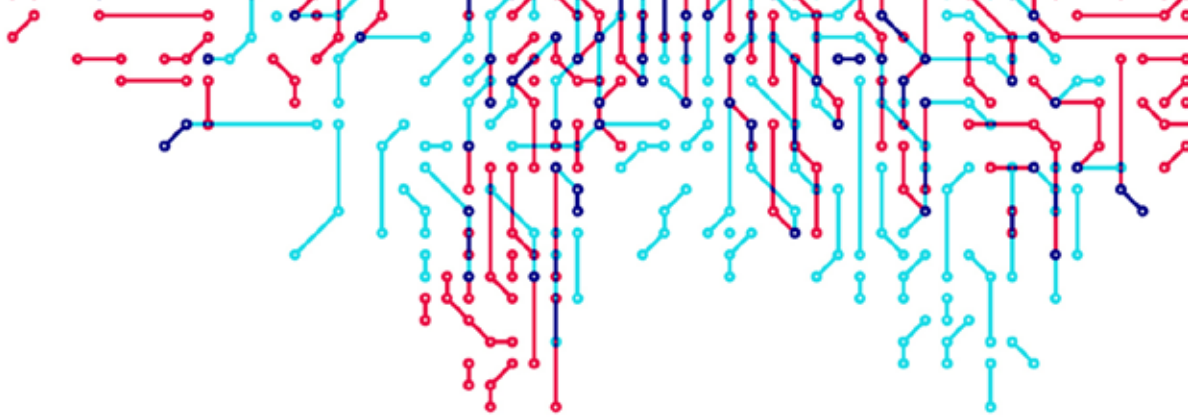
A striking difference between Europe and the United States has to do with opt-in versus opt-out consent for the use of personal data. This means that in Europe consent has to be obtained in an explicit, informed and purpose-specific way. In the United States, consent is tacit: if a person does not want an organisation to process their data, they have to specifically ask it not to.

It could be said that the United States uses artificial intelligence pragmatically by leveraging all the information available.

The relationship between personal data protection and artificial intelligence is therefore an issue which can be addressed in a number of ways. It is up to each society to set the limits and stipulate what is ethical. However, this is not possible if society is unaware of what it is used for and the consequences it has for people. The following chapters seek to answer these questions.







2.
ADA IN ACTION:
RESEARCH FIELDWORK IN CATALONIA

2.1. How artificial intelligence is changing the world

In 2011, the popular American TV programme *Jeopardy*, a general culture-based quiz show, was racking up a $5 million jackpot. Two of the best players in decades, Ken Jennings (74-time winner) and Brad Rutter (who had built up winnings coming to over $3 million), were competing for the big prize. Both accepted the challenge of going head-to-head this time with a very atypical opponent who would be in a room next to the set due to the noise of the ventilation system running it. It was IBM's Watson smart system,[9] a machine capable of understanding complex questions, responding in real time and learning from each interaction. Can you guess who won?

Put simply, you could say that artificial intelligence (AI) is a part of computing which enables machines to operate and react like humans. In other words, they can reason, learn and act intelligently. To achieve this goal, AI develops algorithms which predict and make automated decisions.

The idea is not new. Philosopher and writer Ramon Llull[10] (1232-1316) devoted his life to the *ars machina*, a machine that was used to perform logical tests and make reasoning easier. More recently, mathematician Alan Turing created the *Turing test*[11] (1950) which raised the possibility that a computer could "think". This was demonstrated when a person was unable to decide whether the thing speaking to them was a machine or a human. Turing is also credited with saving millions of lives by breaking the mathematical codes of the Nazi Enigma message encryption machine,[12] which meant he shortened the Second World War by two years.

There are many other historical milestones which have brought artificial intelligence closer to the present, although until now they were solely exceptional cases. It is only today and without us realising it that it is nearer than ever.

Algorithms which help us

Algorithms are the set of commands to follow in order to solve a problem or a task. They were used in humankind's first civilizations, and we ourselves mentally employ algorithms in our brains when we react to an unforeseen situation to resolve it. We do the same thing when we cook a dish, following the cooking time of each ingredient in the recipe.

The algorithms in computers make our lives easier and simpler. When we look for the best route to get somewhere, when we order a taxi on a mobile app, when we get other

9 "Watson" (https://en.wikipedia.org/wiki/Watson_(computer))
10 Ramon Llull (https://www.cccb.org/es/exposiciones/ficha/la-maquina-de-pensar/223672).
11 "Turing test" (https://en.wikipedia.org/wiki/Turing_test).
12 "Enigma" (https://en.wikipedia.org/wiki/Enigma_machine).

recommendations after buying a book, when we watch our favourite series on an online platform, when we look for a house in a new neighbourhood, when we get a job offer, when we ask the bank for a loan, when we get on a waiting list for surgery, when we order a pizza at home or when we take out home insurance, the algorithms decide and show us the best solutions or options. What would we do without them? Giving them up would be a huge waste of time and efficiency!

The most sophisticated algorithms use machine learning, a branch of artificial intelligence which enables machines to improve with experience. It is excellent for establishing huge patterns and relationships and also for streamlining processes. Machine learning includes deep learning, a data processing technique based on artificial, multi-layered neural networks. This technique is inspired by the basic functioning of the brain's neurons. It has been around for more than fifty years but we now have enough data and computing power to apply it to a host of real-life cases.

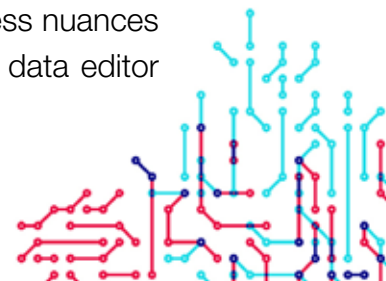
Tireless data manufacturers

The 4.5 billion people connected online today (58% of the world's population) generate an enormous amount of data. Every move we make is already practically a piece of data. We produce data 24 hours a day, 365 days a year. Even when we sleep we are creating data: the data of rest, of non-activity. Without data, the hyper-connected world would slow down so much that it would seem to have stopped.

Our actions on computers and the mobile phones which we hold in our hands all the time produce data that describe our tastes, moods, reactions to stimuli, fears and ambitions. The most valuable data for companies, political parties and governments are any which define the behaviour of groups of people. After analysing and classifying them, they profile purchasing patterns to send us product and service offers or messages which are so personalised (and so much to our liking) that we can hardly reject them. Consultancy firm Cambridge Analytica used these performance analysis techniques to spread misinformation (popularly called 'fake news') through Facebook in political campaigns such as the ones run for President Donald Trump and Brexit.

There are other data including fingerprints, the face, the iris or the genome which identify us as unique and unrepeatable beings. There are also accumulated data derived from our past actions which classify us as good or bad payers or which open and close the doors to new jobs depending on where we live, our gender, age or skin colour.

Nowadays data decide everything about our lives in the present and in the future. It is as if we had suddenly put on prescription eyewear and discovered countless nuances in our environment which we had never noticed before. The Economist's data editor



Kenneth Cukier[13] explains in his book *Big Data: A Revolution That Will Transform How We Work, Live and Think* that the value of information today lies in how "we can discover patterns and correlations in the data that offer us novel and valuable insights."

We trust to evolve

We already accept it as normal that a machine or virtual assistant bids us good morning every day, we have a device which goes around the house cleaning or that an automated system suggests the series we want to watch based on what time it is. Soon we will accept that the fridge does the week's shopping, the washing machine takes care of the laundry independently and that the house's heating calls the engineer when the boiler breaks down. All these objects generate (or will generate) data related to consumption, interactions and use. It is known as the Internet of Things (IoT), i.e. the digital interconnection of objects to communicate and interact with each other. It is estimated that by 2025, there will be 21.5 billion connected objects in the world.[14] We will also have domestic robots which will take care of the elderly and children at home, students in class will study the subjects most suited to their personality and we will trust autonomous cars as the safest vehicles for our journeys.

We will trust technology because we know that otherwise we will not evolve as a society. We will trust it because history has shown us that technology has always moved us forward as human beings.

Complexity can be understood with examples

Yet to trust it, we have to know the risks and rewards of the technology around us and the technology we are promised. Only in this way can we adopt a critical mindset and assess how much technology we want in our lives.

The purpose of this research is to kindle individual curiosity, to explain in a simple way what the automated decision-making algorithms (ADAs) already in use in Catalonia are for. To explain their complexity, we have compiled more than fifty examples in social, employment, healthcare, education, banking, media and communication, cybersecurity and other fields.

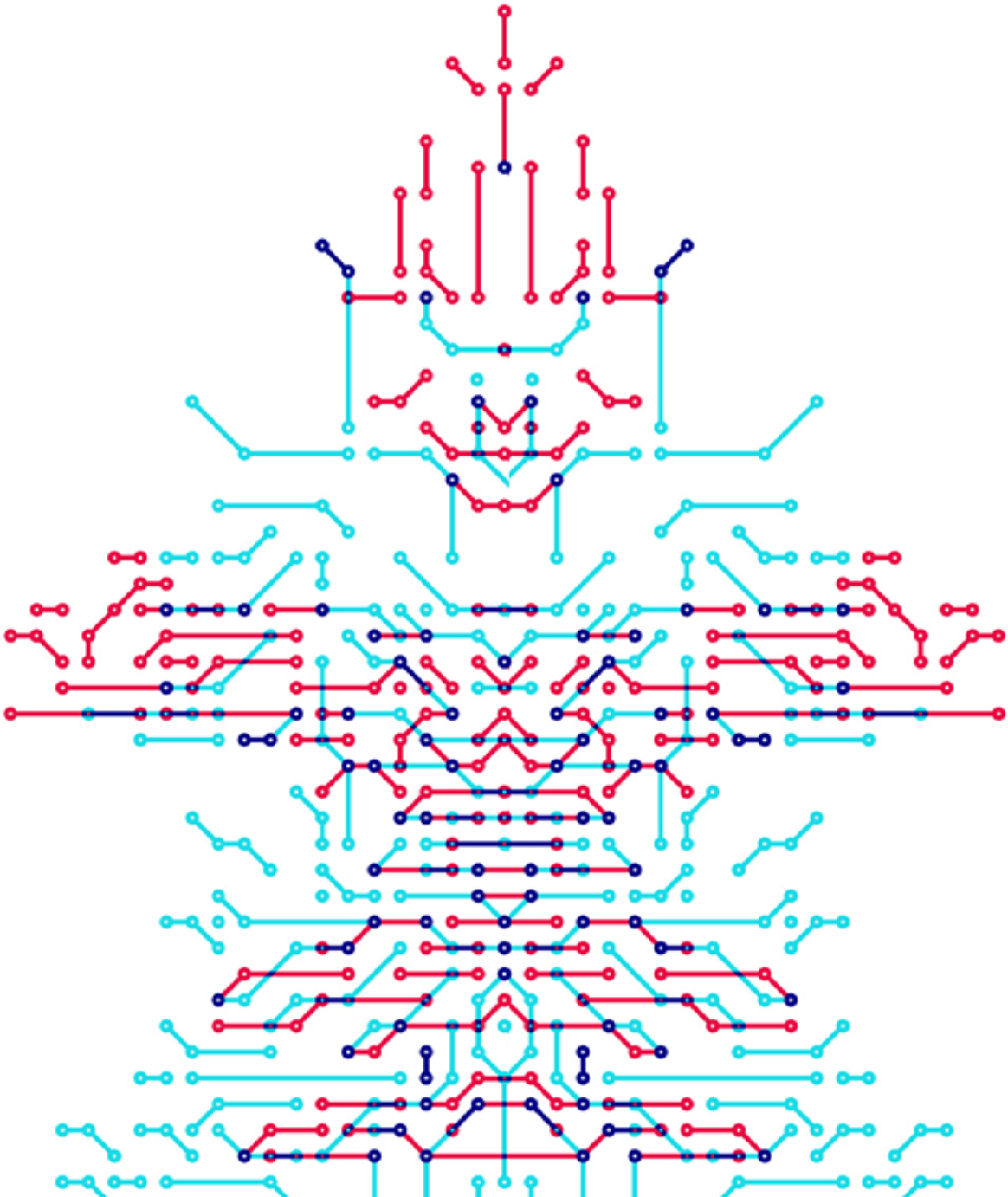
In the preparation of this study, we have benefited from the selfless cooperation of some thirty experts in Catalonia, the leaders in academic research in artificial intelligence, and

13 Kenneth Cukier (https://www.ted.com/speakers/kenneth_cukier).

14 Report *La Internet de les coses a Catalunya*. Acció. Generalitat de Catalunya (https://www.accio.gencat.cat/web/.content/bancconeixement/documents/pindoles/iot-cat.pdf).

of entrepreneurs and professionals involved in AI, who have shared their knowledge and thoughts on current developments and the pitfalls to bear in mind. Undoubtedly we have not mentioned a lot of very useful research and initiatives in the abovementioned or other areas. This is just a sample to help us understand and decide.

Because that is what it is all about: deciding individually what to do if an automated decision-making algorithm adversely affects us or discriminates against us. To know which agencies or legislation can protect us. To know the procedures to follow so that the criteria used to design the algorithm are reviewed, as well as to identify and mitigate the biases in the data which have led to an unwanted effect for a person or a group.



2.2. Automated decision-making algorithm risks

Imagine you are looking for a job at an extremely respected company offering more than acceptable working conditions and salary. The company you want to join receives many applications and does not conduct personalised interviews or tests because it believes that the skills shown in CVs are often overstated. Instead, it asks for the candidate's email password so that an algorithm can scour their personal messages and decide whether they are the person the company is looking for. Would you accept having your mailbox checked by an artificial intelligence system in exchange for the chance of getting the job of your life?

The example is real and is described in the report prepared by the NGO Algorithm Watch, *Automating Society: Taking Stock of Automated Decision-Making in the EU*.[15] The Finnish recruitment agency DigitalMinds has about twenty large corporations as clients. Since it receives thousands of CVs every month, it can only select candidates by using automated decision-making algorithms. DigitalMinds also uses them to scan applicants' personal Facebook and Twitter accounts. The system analyses the candidate's activity and how they react. The results can show whether a person is introverted along with other personality aspects. Again the question: would you accept this? As the technology already exists, we should not rule out similar methods soon being used in Catalonia.

They take decisions for humans

Algorithms today are already capable of taking over some functions from humans and this means they are increasingly essential. They can do facial or image recognition, interact (virtual assistants), decide automatically (series, book or other product recommenders), have a social impact (robots which look after patients in hospitals), and learn from many situations in real time (the autonomous cars we will soon have).

"There are three types of automated decision-making algorithms," notes Jordi Vitrià, Professor of Languages and Computer Systems at the University of Barcelona (UB). "Ones that predict a number (I'll make 33 cars); ones that predict a class (true or false, yes/no, ill/healthy); and the recommenders, which out of a very large set of options ensure that you'll buy a certain product. These three types first make the prediction and, depending on the result, take the decision (automated or otherwise). "

Elisabeth Golobardes has a PhD in computer engineering. "There are predictive and automated decision-making algorithms," she says. "They can predict leads in a business setting or a potential accident in a nuclear power plant. And then there are ones that

15 Report drawn up by Algorithm Watch, *Automating Society: Taking Stock of Automated Decision-Making in the EU* (https://algorithmwatch.org/en/automating-society/). Karma Peiró wrote the "Spain" chapter.

make decisions such as evacuating the at-risk population 100 kilometres away from the core of the nuclear power plant due to a possible explosion, or deciding to lay off a hundred workers at a company owing to likely bankruptcy. These are very complicated decisions and I hope that behind them there will always be a professional who will assess the situation and be accountable if the machine's decision is implemented in an automated way."

Traditionally banks have used many algorithms before giving or refusing a loan. The only difference with today is that previously it was the bank manager who decided based on variables shaped by years of experience, whereas now the process is automated by machines. It is the same with car or home insurance: what a professional, the traditional salesperson, used to do is now done by a machine. It is all about having a large database with thousands or millions of cases from the past and training the algorithm. "But beware of biases!" warns Vitrià. "If I take a database from a bank and for thirty years there has been a male chauvinist view of society whereby they didn't give loans to women, the algorithm will perpetuate this bias and today they won't give them either." Algorithms are neither good nor bad, but they are not neutral either. They are fed by data, and the data have biases. "They have always existed," he adds. "What we have to do is identify them and mitigate them."

Damn biases!

In the 1980s, Dr Geoffrey Franglen, Deputy Dean of St. George's Hospital Medical School in London, had to evaluate some 2,500 applications each year. To automate the process, he wrote an algorithm to help him review them based on the evaluation behaviour of previous applications. That year, the candidates underwent a double test before being admitted: the algorithm test and the teacher test. Franglen reported that the scores were 90-95% consistent, showing that AI could replace humans in this very tedious phase. But four years later, the school's management realised that there was little diversity in the candidates. And the UK Commission for Racial Equality reported the school for xenophobic and gender discrimination. It turned out that every year the algorithm had left about 60 people out of the selection process: it discriminated against them because they had non-European surnames or because they were women. Biases were being perpetuated.

"There are three types of classic biases: statistical, cultural and cognitive," says Ricardo Baeza-Yates, Professor of Computer Science at Pompeu Fabra University and Northeast-

ern University.[16] "Statistical bias comes from how we get the data, from measurement errors or the like. For example, if there is a greater police presence in some districts than in others, it is not uncommon for the crime rate to be higher where there are more officers. Cultural bias stems from society, from the language we use or from everything we have learned throughout our lives. The stereotypes of people from a country are a prime example. Finally, cognitive bias is what identifies us and depends on our personality, tastes and fears. If we read a news item that is in line with what we think, our tendency will be to accept it even if it is fake."

This latter factor is also called *confirmation bias*. A lot of fake news is driven by this way of thinking so it spreads more quickly. Hence if we do not question what we read or see, we run the risk of going backwards as humans. Historian Yuval Noah Harari[17] warns in his latest book *21 Lessons for the 21st Century* that "with today's technology, it is very easy to manipulate the masses". And if we follow what most people think, what happens when the masses are morally wrong?

And even more biases...

Ricardo Baeza-Yates says[18] there is a ranking bias when we search online as people tend to click the top positions and the search engine might interpret this as meaning that these responses are better than the lower down options.

Presentation biases are the ones we find in recommendations in ecommerce. Only what the search engine shows the user can get clicks. Anything which does not appear on the results page is excluded from the query. It is a vicious circle, like the chicken and the egg. And the only way to turn it around is to show the entire universe of results. "This is known as a filter bubble: the system only shows you what you like," adds Baeza-Yates. "As it is based on past actions, it is not possible to see the unknown."

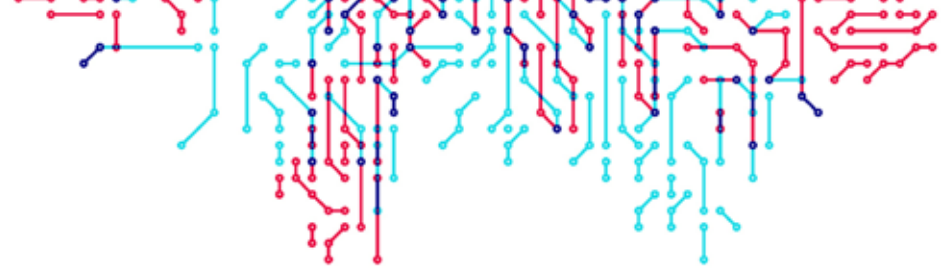
If the world continues to work in this way, there will come a time when we will feel like the main character in the movie *The Truman Show,*[19] who one day realised that his whole world was a hoax and he had lost what is over the horizon. Social media work like this and multinationals are fine with it. They promote what is called the dopamine economy. Dopamine is a neurotransmitter we have in our brains which provides a sense of pleasure, feelings of joy or reinforcement to motivate people to take action. Being in the

16 Ricardo Baeza-Yates, Karma Peiró. "És possible acabar amb els biaixos dels algorismes (1a i 2a part)". 2019 (https://www.karmapeiro.com/2019/06/17/es-possible-acabar-amb-els-biaixos-dels-algoritmes-1a-part/)

17 Yuval Noah Harari (https://www.ynharari.com/).

18 'És possible acabar amb els biaixos dels algorismes (1a i 2a part)', Ricardo Baeza-Yates and Karma Peiró (2019) (https://www.karmapeiro.com/2019/06/17/es-possible-acabar-amb-els-biaixos-dels-algoritmes-1a-part/)

19 Movie *The Truman Show* (https://en.wikipedia.org/wiki/The_Truman_Show).



filter bubble of a social media site can create an addiction which prompts us to spend countless hours interacting.

Biases that discriminate against women

In 2014, Amazon unveiled an algorithm to recruit new workers for its warehouses.[20] The tool rated the best candidates from one to five stars. Everything seemed just right: AI would save hours in the human resources department. Yet a year later, the multinational realised that no women had been hired in technical positions such as software developer. Could it be that there were no candidates with the right skills for this job?

The example is already a classic of the errors which AI has recently engendered. The big data which fed the recruitment algorithm were based on CVs received in the previous decade in which most of the programmers were men. When the automatic system detected the word *woman* or a synonym for it, it directly penalised the CV by giving it a lower score.

Another example of gender discrimination took place in 2016. Researchers from Microsoft Research and Boston University used mass collection of Google news to train algorithms on male/female stereotypes in the press. The results showed that men were computer programmers and women were housewives; men were doctors and women were nurses. And it makes sense, because in the United States most of the journalists who wrote the news on which the training algorithm was based were men. So Google simply reflected the gender bias that actually existed.

There are loads of kinds of bias. There are dozens of cultural examples and even more cognitive ones. Up to 100 have been classified, but about 25 are the most important. The article "The Ultimate List of Cognitive Biases: Why Humans Make Irrational Decisions"[21] lists 49 cognitive biases. These are the most dangerous because they are ingrained in every person. The only way to overcome them is to change each person, which at first sight seems an impossible feat. The historian Harari is right when he says that "it is very easy to manipulate people and very challenging to eliminate biases."

Yet smart systems are not autonomous and nor do they act on their own. Researcher Elisabeth Golobardes explains this very well: "The algorithm as such has no bias. It is the data you enter and the purpose for which it has been designed that discriminates."

20 "Por qué la inteligencia artificial discrimina a las mujeres?". Ricardo Baeza-Yates, Karma Peiró (2019) (https://medium.com/think-by-shifta/por-qu%C3%A9-la-inteligencia-artificial-discrimina-a-las-mujeres-18b123ecca4c).

21 *The Ultimate List of Cognitive Biases: Why Humans Make Irrational Decisions* (https://humanhow.com/en/list-of-cognitive-biases-with-examples/).

Can you be fair to everyone?

Not all biases are harmful either. “For example, having more female nurses than male nurses may be a good thing because of their empathetic bedside manner,” says Baeza-Yates. “By contrast, the fact that male politicians are in the majority is not because one point of view of the population (the female one) is not equally represented.” Algorithm results can discriminate on the basis of gender, race, age or social class, just to cite the most important factors.

Knowing that algorithms have biases and may discriminate, then why are they used? “One answer might be that the reward or accuracy of the results is considerably higher (over 90% in most cases) than the harm or error,” continues Baeza-Yates. “Is this fair to the people who are harmed? At this point you could start a long discussion about what is and is not fair in life.”

It is very difficult to be fair to everyone. An algorithm can be fair to a group of women yet discriminate against a man. Andrew Selbst,[22] a researcher at the Data & Society Research Institute,[23] explains that discrimination in artificial intelligence is very complicated: “It is a constantly evolving process, just like any other aspect of society.”

To be even-handed, it should also be said that well-designed algorithms, even when they have biases, are fair according to the parameters they have been given. Unlike humans, who can vary their decisions depending on mood or physical and mental fatigue, algorithms always work in the same way.

Like prejudices

Biases are similar to prejudice: we all have them to some extent or another. Many of us inherit them from our social or family environment without realising it. The biggest bias is believing we don’t have any prejudices. But watch out! If biases are not remedied, there is a risk we will live in a future in which social progress will be increasingly tricky because prejudices have been perpetuated.

So how can we be sure that all the data we pump into the algorithm represent the universe we want to predict and that they will not discriminate against anyone? We can’t. And here we have the ethical aspects of artificial intelligence which we need to keep in mind. (See chapter *6, “Ethical aspects of AI with the views of researchers”.*)

22 Andrew Selbst (https://andrewselbst.com/).
23 Data & Society Institute (https://datasociety.net/).

2.3. Automated decision-making or support for the decision?

On one of her visits to Barcelona, mathematician Cathy O'Neil,[24] author of the book *Weapons of Math Destruction* which highlights automated decision-making algorithms (ADAs) in the United States, posed this dilemma for me: "Would you be able to decide whether a child is in danger at home because they are about to be sexually mistreated or abused by one of their parents?" Seeing my concerned expression, she added: "In the United States, it's the machines that are deciding. The Aura algorithm[25] was built to identify potential victims of abuse before the event with the good intention of avoiding trauma for the child." Puzzled and quite concerned, I asked her how a machine can know for sure that a child should be taken away from their parents before anything happens. And O'Neil's reply was: "It can't." The Aura algorithm was a pilot scheme that in the end was never implemented. However, there are other automated systems such as COMPAS[26] which predicts crime in prisons that have been extremely contentious. Investigation by the independent media outlet *ProPublica*[27] demonstrated it was biased against people of colour.

We have not gone this far yet in Catalonia. We are at an early stage of artificial intelligence and many of the examples we have are still part of research or pilot trials. Furthermore, all the experts we have spoken to for this report make it clear that automated decision-making is only used in sectors which do not involve a life-threatening risk for people. So in healthcare, the judicial system and in some cases in education, the algorithm's decision is a support to the professional and is not applied in an automated way. The last word is always left to the doctor, judge or teacher concerned.

A matter of scale

In areas such as job recruitment, where demand is in the hundreds or thousands of people, the algorithm draws up a shortlist. When awarding social benefits with many applicants in the same conditions, the automated system decides. And in some ways it can also lead to major changes and adverse consequences for people's lives.

"Automating staff selection in human resources departments is a very common example," says Ramón López de Mántaras, research professor at the Spanish National Research

24 Cathy O'Neil (https://mathbabe.org/about/).
25 "Cathy O'Neil, author of 'Weapons of Math Destruction,' on the dark side of big data". *Los Angeles Times* (https://www.latimes.com/books/jacketcopy/la-ca-jc-cathy-oneil-20161229-story.html).
26 "Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)" (https://en.wikipedia.org/wiki/COMPAS_(software)).
27 "Machine Bias", *ProPublica* (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Council (CSIC). “If you get a thousand CVs every month, or every week, you can’t look at all of them and make a choice. The algorithms pre-select. And managers only look at the ones that have got through the ADA filter. But machines don’t see things that humans do. And that’s when the discrimination begins.”

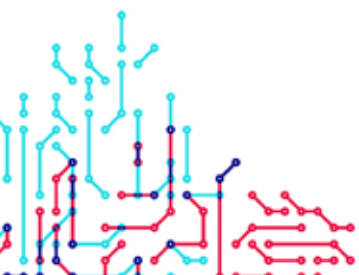
“It’s different in medicine,” he adds. “Doctors only use ADAs as decision support. The Intensive Care Unit is jam-packed with algorithms and obviously one or another of them might trigger an alarm, but a doctor (or healthcare worker) looks at what’s going on and makes the final decision. There is no automated system which will administer a drug without professional supervision. And I hope it will always stay that way because the doctor sees things the algorithm doesn’t. Then again the algorithm finds patterns which the human eye can’t make out. The algorithm looks at the trees and the doctor looks at the wood. So when they work together it greatly reduces error. In breast cancer screening, studies show that the best doctor has a 5% or 6% error with mammograms and an ADA 6% or 7%, but that working together the error is only 0.5%.”

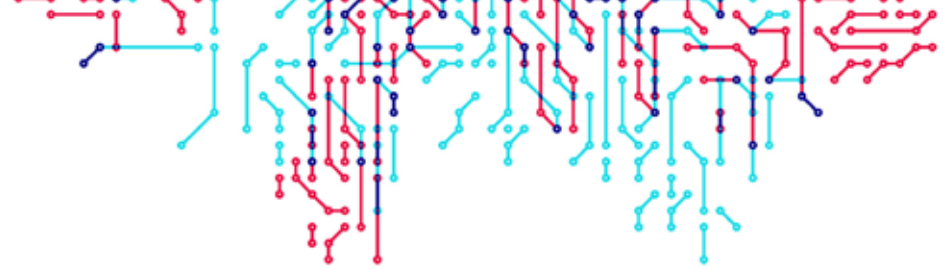
Black boxes

Let’s recap: artificial intelligence uses algorithms that learn from big data. These data have biases which can discriminate. They have to be identified and mitigated. Algorithms predict and then some of them make automated decisions. A substantial number of these systems are machine learning, i.e. they learn before they are used, while others are deep learning.

Any system that is considered intelligent must have the ability to learn, i.e. to improve with experience. What has not yet been explained is how the algorithm came to the prediction or decision. This is called “black boxes”. Applications for natural language processing such as translators, medical diagnosis, bioinformatics and identifying financial fraud are black boxes.

So if an automated system predicts an action that discriminates against me, how did it come to that decision? We can ask the question but for the time being we will not get an answer. And this is currently one of the great dilemmas in AI.





Explainability or algorithmic transparency

Understanding how the algorithm made the prediction or took the decision involves explainable artificial intelligence.[28] "A lot of work has been done on this problem in machine learning algorithms over the last two years," says Baeza-Yates. "In particular about the most opaque, which are the deep learning ones, but we still don't have all the answers."

Oxford Institute researcher Sandra Wachter[29] believes that we should have a legal right to know why algorithms make the decisions that affect us. She says that the owners of the algorithms - multinationals, companies of all sizes, banks, and also governments, public agencies and the police - will go to great lengths to avoid making the formulas of their systems transparent in order to safeguard intellectual property or public safety. That is why she suggests that "counterfactual explanations" should be given. In other words, if your mortgage application has been rejected, you can ask the bank: "If I earned €10,000 more per year, would you have given it to me?" If you have failed to get a job, you can ask: "If I'd had a master's degree, would you have hired me?"

The scientific community is striving to get algorithms which provide explanations of their decisions. And we might think that a way will be found very soon. We are at the beginning of a social change which will only get bigger in the coming years and the public has a very important role to play in it. Technology has its bright and dark sides. The dark sides are inadvertent errors but also intentional misuse with respect to a person or a group depending on the biases of the data or the criteria used to build the smart system. We need to be on our guard against anomalies and discrimination in order to ensure through legislation and the public authorities that privacy, confidentiality and individual freedom are still safeguarded.

28 "Explainable artificial intelligence" (https://en.wikipedia.org/wiki/Explainable_artificial_intelligence)

29 Sandra Wachter. "How to make algorithms fair when you don't know what they're doing". *Wired*. (https://www.wired.co.uk/article/ai-bias-black-box-sandra-wachter).

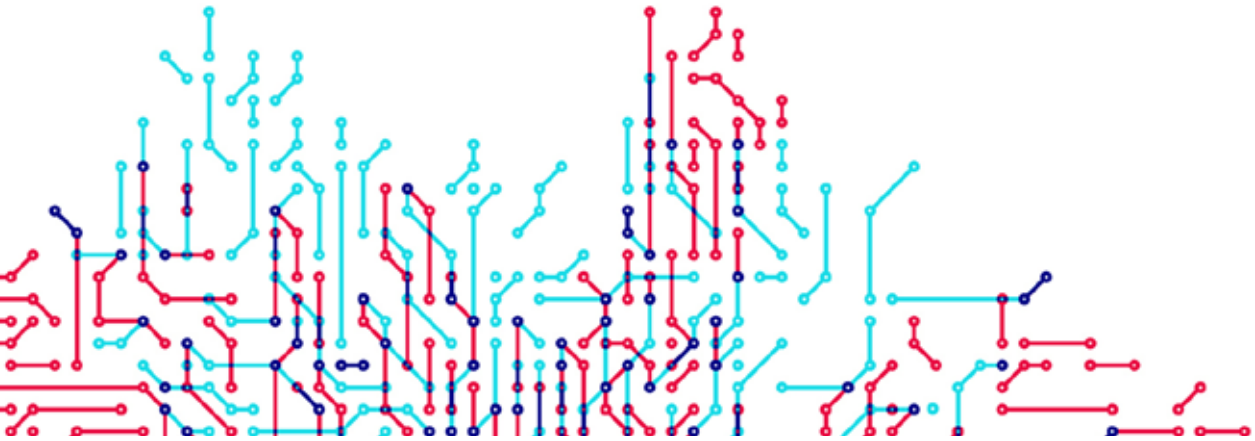
ADA IN ACTION

2.4. Where are ADAs used in Catalonia? Over 50 examples to help you understand…

The best way to understand technological concepts is with examples. When we put ourselves in a certain situation, we can picture more clearly what the machine in question would do or how the smart system would act. Only in this way can we assess the importance, risks and consequences of their use.

So we thought it was crucial to include in this report a section setting out where automated decision-making algorithms (ADAs) are used in a range of fields: healthcare, the judicial system, education, social issues, business, cybersecurity, banking, employment, the media and communication, etc. This is by no means an exhaustive compilation and we are fully aware that we have left out an uncountable number of examples. But we do think it is a sample which gives an idea of how invisible ADAs are and helps to understand both the value of their use and the risks involved.

Generally speaking, it is fair to say that while their implementation in business and the private sector is well advanced, there is still a long way to go for government. It is also significant that the potential afforded by artificial intelligence to move forward towards a more efficient and fairer world have been explored for years in the healthcare and judicial sectors in partnerships between local and international hospitals, universities, private organisations and start-ups or with proprietary developments.



2.4.1. Healthcare

Artificial intelligence is a great support in healthcare. Algorithms help to understand or draw conclusions about diseases much more quickly, suggest a relevant diagnosis and medication, better manage hospitals and read medical records on a large scale.

In this case and unlike in other sectors, automated decision-making does not take place without the supervision of a doctor or specialist in the field. And this is important. What would happen if the algorithm failed to detect cancer in a patient and the doctor did not provide any treatment? In Catalonia, at least at the time of writing, AI is always a support for the decision of the specialist doctor.

Artificial intelligence has enormous possibilities but always coupled with the potential and knowledge that healthcare professionals have. As noted in the previous section, when the automated decision involves a life-threatening risk, the specialist always has the last word.

"The human brain is very slow compared to what machines can do today," says researcher Itziar de Lecuona, a lecturer in the Department of Medicine and Deputy Director of the Bioethics and Law Observatory at the University of Barcelona. "But this doesn't mean we should blindly trust technology. All models have to be validated, proofs of concept have to be carried out and we need to ask ourselves whether the algorithm is working towards the initial objective."

"There have always been algorithms in medicine, but now we've taken another step: we're moving towards prediction, towards a more personalised experience. With today's technology and its capabilities, it would be very inefficient not to use it." Lecuona says that in healthcare everyone always thinks about doing good: "But ethics also need to be borne in mind because algorithms are fed with data that are part of people's privacy."

Meanwhile Dimitra Liveri, a Security and Resilience of Communication Networks Expert at the European Network and Information Security Agency (ENISA), spoke at the Barcelona Cybersecurity Congress[30] about smart hospitals, which will soon no longer be the exception anywhere. "A smart hospital is an ecosystem of interconnected machines which make automated decisions," she said. What would happen if a smart insulin pump made a mistake? The patient might die. The ENISA expert noted in her presentation that there are still challenges to be addressed. "1. The availability of the device: if I'm ill I want my device to work all the time non-stop. 2. Integrity: I want it to inject me with the right doses of glucose and never make a mistake."

30 Barcelona Cybersecurity Congress (https://www.barcelonacybersecuritycongress.com/es/front-page/).

• A pill that videos the intestines

Colon cancer in Catalonia is a disease that mainly affects people over 50. It is estimated that about 5,000 people are diagnosed each year and it is associated with their lifestyle and diet. For years the Government of Catalonia has been working on getting tests to detect it at home. Using a simple test, it can be identified in time in case of anomaly. When the test is positive, an endoscopic test or colonoscopy is performed, which in some cases requires hospital admission and sedation and is costly for the public health-care system.

In the USA and in some European countries a new technique is being used that might be implemented in Catalonia before long. Researchers at Hospital Vall d'Hebron together with the Department of Mathematics and Computer Science at the University of Barcelona (UB) are experimenting with an endoscopic capsule[31] to study the condition of the intestines. Patients swallow a small pill which is fitted with a camera, four LEDs[32] and a battery. The pill travels through the patient's body and records images that can be extremely helpful to the doctor. The images are sent via Wi-Fi connected to a device worn by the person on a belt. The recording takes between eight and twelve hours.

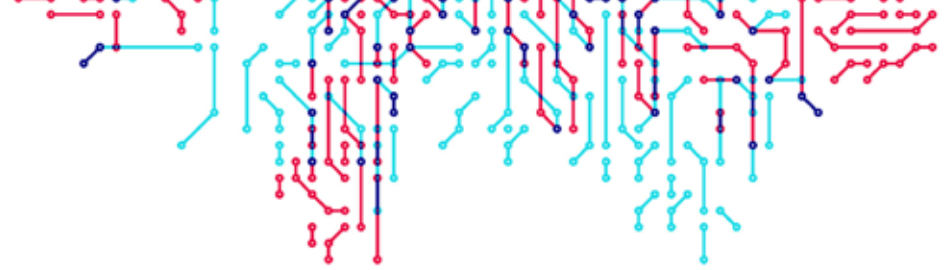
The pill has no contraindications and if it raises any concerns the colonoscopy is performed. The invention seems just the thing, but then the question arises whether doctors have time to watch so many hours of recording of each patient. "Impossible," replies Jordi Vitrià, Director of the Department of Mathematics and Computer Science at the UB. "Either you make an automated system that analyses and detects whether it sees any anomalies or it doesn't make sense. Depending on what the algorithm detects, the doctor will decide whether to perform an endoscopy or other types of tests."

• Exaggerated pain detector

Let's imagine that a worker has had an accident or an injury to an arm, a leg or their back and asks for mid/long-term sick leave. The length and amount of the benefit depends on how long they have to be off work. As a rule these cases are handled by occupational insurers using a physical examination in which the injured worker is put through various machines which test out their joints. But there are always a percentage of cases in which the insurers cannot really decide on how bad the pain is.

"It has been proven that a high percentage of people exaggerate their pain," explains Vitrià. In other words, the part of the body doesn't hurt as much as they say it does. When this is found to be the case, the Social Security authorities deny the sick leave.

31 "Endoscopic capsule" (https://hospital.vallhebron.com/en/diagnostic-tests/endoscopy-endoscopic-capsule).
32 "LEDs, miniaturised light-emitting devices" (https://en.wikipedia.org/wiki/Light-emitting_diode).



"There is a lot of public money at stake," says Vitrià. "That's why they asked the UB to create an algorithmic model based on the whole history of tests and diagnoses built up over the years. Now we can see that in many cases the patients are right, but in others they aren't. Pain exaggeration can be detected in an automated way." After being introduced in Barcelona, the system has been implemented in Seville with the same criteria so that pain tolerance levels and the decisions taken are similar in both cities.

• Smartphones for the elderly

The number of people over 80 has increased in Europe in recent decades. Higher life expectancy coupled with a declining birth rate should lead us to consider how to invest resources more efficiently. Longer life does not necessarily mean better health. What can be done to maintain the wellbeing of the elderly? What fine-tuning is needed in social, economic and public health policies?

The NESTORE European project[33] is a virtual coach that supports the elderly who are in good health and gives them personalised advice about their eating habits and the daily exercises they need to do. Tracking is done through connected objects and sensors spread around the house and a mobile app which monitors all their movements.

"They are given a smartphone and they have to photograph what they have in the fridge, what they eat, etc.," says Itziar de Lecuona, a NESTORE researcher. "Computer vision algorithms are used to predict the intake they need based on their state of health. In the home they also have a device (a virtual assistant) which reminds them that they have to go out for walks, the minimum number of steps to take or other types of physical exercise. They end up becoming a very close companion because they know everything about you: illnesses and medication, food, hours of sleep, whether you have been active or not, whether you have been in contact with your family, whether you have had visitors, etc."

• Matching medication after a transplant

Hospital del Mar in Barcelona has one of the largest kidney transplant units in Spain. It started out in 1973 and today has already performed more than 1,400 transplants.[34] In Catalonia, almost half of the new patients undergoing kidney replacement therapy are over 70 and a large number of kidney donors are over 60.

One of the main problems after a transplant is getting the right lifetime dose of medication given to the kidney recipient. This means that this step is very delicate. "If you overdo it,

33 NESTORE European Project (https://nestore-coach.eu/home).
34 "40 anys de trasplantament renal a l'Hospital del Mar" (https://www.parcdesalutmar.cat/es/noticies/view.php?ID=1003).

it may have side effects, and if the medication is too low, the organ may be rejected by the body. It is very tricky," explains Fernando Cucchietti, Director of Data Analysis and Visualisation at the Barcelona Supercomputing Centre (BSC). "We suggested Hospital del Mar should use algorithms to decide whether or not a person receives a kidney with all the factors that are currently considered between donor and recipient, but also to help them specify the exact medication protocol." The project has not yet been started up.

• Early cirrhosis diagnosis

Whenever the liver is injured, whether by disease, excessive alcohol consumption or another cause, the organ tries to repair itself. In the process, scar tissue forms. As the cirrhosis progresses, more scar tissue forms and this makes it hard for the liver to function (decompensated cirrhosis). Advanced cirrhosis is life-threatening. However, if the cause is diagnosed early and treated, the damage can be limited and sometimes reversed.

Hospital Clínic de Barcelona has one of the best liver treatment units, yet by the time patients come in it is too late. Cucchietti points out that this is why it's so important to develop early detection technologies and systems. "We are now part of a European project with Hospital Clínic which will allow us to get the preliminary symptoms of liver cirrhosis," he says. "The trial will be carried out on 40,000 people across Europe over five years and is coordinated from Barcelona." The idea is to install a device in primary care centres and hospitals so that the algorithms make the diagnosis and doctors decide whether the patient should have a biopsy or a transplant. "You want to detect the symptoms five or ten years before it's too late, because in this early stage you need a change in lifestyle and very little medication."

• Diagnosis by digital electrocardiograms

A vascular health company manufactures devices that perform digital electrocardiograms. This allows them to be processed and associated with diseases. "Having all this knowledge encapsulated is like having a committee of doctors who over time have diagnosed a series of diseases with many hundreds of thousands of electrocardiograms," explains Jordi Navarro, a data expert and CEO of a Catalan company engaged in AI.

"The algorithm makes a diagnosis similar to the doctor's based on finding patterns. The difference is that if there were 20 cardiologists, one would be wrong in the diagnosis and 19 would be right, but the algorithm always takes the majority's result; in this case, the right one. So in this respect we might say that the algorithms are democratic."

• Facial recognition to identify ADHD

Attention-deficit hyperactivity disorder (ADHD) is neurobiological in origin and consists of the inability to maintain attention along with hyperactivity and impulsivity. It has a significant impact on a child's life. The symptoms of ADHD may appear at school or in any other social setting. The disorder begins in childhood but can continue into adolescence and adulthood. It is estimated that between 3% and 7% of children may be affected worldwide. In Catalonia, 1.3% of girls and 3.7% of boys aged 6 to 17 resident in the region took some kind of drug for ADHD during 2015. This figure is 1.9% and 5.5% in girls and boys aged 12 to 15 respectively, which is the group with the highest usage.[35]

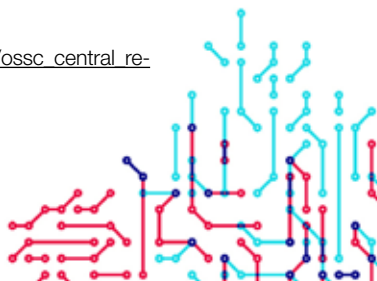
Facial recognition can now help with more accurately diagnosing the disease. The Computer Vision Centre (CVC) at the Autonomous University of Barcelona (UAB) has started up a smart system which analyses the child's gestures and facial expressions. The results can be used to assess the symptoms with a degree of confidence. This diagnosis is used by the doctor/psychologist who is treating the child to carry out the action protocol. The same algorithm is also used to try to diagnose other depressive diseases among young people. "It's just starting out," explains Meritxell Bassolas, Director of Knowledge and Technology Transfer at the CVC. "In addition to the results of the algorithm, the child's behaviour on social media is analysed. All this means the healthcare professional has more information to apply a treatment." *(See section 2.4.10 for other examples involving computer vision.)*

• Patients don't go to the hospital assigned to them

CatSalut is the agency which manages the public health system in Catalonia. It allocates healthcare resources among counties in Catalonia, which are grouped into nine health regions and into sub-regions of various categories. Everyone is assigned a primary care centre along with a hospital for emergencies or longer stays. However, people do not always go to the facility they have been assigned, which means facilities initially planned for a certain number of people may be underused or overused. Why do patients not go where they should by place of residence? There may be many reasons: logistical problems, lack of healthcare professionals in the facility, overly long waiting times or inadequate public transport timetables for getting there.

Until now this question was answered by the painstaking work of experts who looked for the anomalies after extensive observation of operations. Now a program can do it

35 Catalan Health System Observatory (http://observatorisalut.gencat.cat/web/.content/minisite/observatorisalut/ossc_central_resultats/informes/fitxers_estatics/MONOGRAFIC_26_TDAH_CdR.pdf).



automatically:[36] the algorithm takes the behavioural variables of the patients and displays them in green or red depending on whether they are going to the right place or not. With this information it extracts anomalous patterns from each health centre. This means they are quicker at spatial organisation and mitigating a problem with public facilities.

• Predicting when a patient will go back to hospital

People with chronic heart or kidney disease go back to hospital at particular intervals. Sometimes this is because of small destabilising events such as a hot day or running to catch a bus. The hospital readmission indicator, which is internationally recognised as a quality test,[37] sets 30 days as the optimal rate for a patient to return to hospital. "If the patient comes back within a month, something has gone wrong," says Ricard Gavaldà, lecturer and coordinator of the research laboratory at the Technical University of Catalonia (UPC). "Maybe the hospital needed beds and the patient was sent home earlier than planned, maybe there has been no communication with the GP in the follow-up of the medication administered, or simply because the person needed more care."

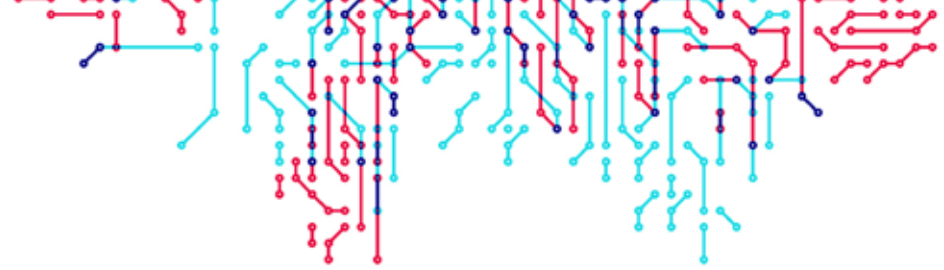
"In Catalonia, CatSalut pays 100% of hospital costs. But if the patient comes back 29 days after leaving, the hospital only gets 40%. The underlying aim is to provide better care," explains Gavaldà. Using the basic information of thousands of cases from a hospital, a program has been built which predicts the risk of the patient coming back before 30 days are up. "When there's a high risk, the hospital can keep the patient in bed for a couple more days. With a low risk, you can do daily follow-up at home to find out whether everything is going okay, but you still have a bed for another emergency," says Gavaldà. "This saves public resources and improves the patient's wellbeing if they don't need to be readmitted to hospital."

Gavaldà has no doubts about the automated decisions that algorithms can already make in healthcare for any disease diagnosis: "It is not a question of whether the machine does it better than the doctor or not, but that the doctor does better with the machine than without it."

• Expediting triage in A&E

When a person goes to A&E at a hospital, especially in big cities, they have to endure hours of waiting in the entrance hall, in a corridor or in a cubicle. After a long while they

36 Amalfi Analytics (https://www.amalfianalytics.com/).
37 "Readmission rate as an indicator of hospital performance. The case of Spain" (https://www.researchgate.net/publication/8267981_Readmission_rate_as_an_indicator_of_hospital_performance_The_case_of_Spain).



are normally seen by a doctor who sends them for tests. Hours later, sometimes a whole day and night, the doctor decides whether to assign them a bed.

What would happen if right from the start in the admission triage a bed was requested for the most recurrent cases? Leveraging the thousands of admissions there have been in A&E and the data on how they have been dealt with, an algorithm can make an automated decision about which people will need one. They would be added to the bed waiting list from the time of triage and not after having spent hours waiting.

If the machine gets it right, the patient only has to wait for the time it takes for the tests ordered for the diagnosis. This means the hospital gains in efficiency. If the algorithm is wrong and the patient does not need the bed, it will quickly be reassigned to another patient.

• Translating medical records

When a researcher wants to conduct a search they find the big problem that 80% or 85% of the records are in what is called *free format*. In other words, the patient explains what is hurting and the doctor adds it to their records. The diagnosis of the disease, the treatment, the medicines administered, etc. all remain in a format that machines cannot read. And with the information in this state, no large-scale investigation can be performed.

Hospital Vall d'Hebron uses a program[38] that automates and converts medical records into structured data for later analysis. By hand this task is very slow and always lags a few years behind. "We use natural language processing," says Gabriel Maeztu, a doctor and data scientist on the project. "In other words, we have taught the computer how to read medical records using the variables asked for by the researcher. The algorithm reads them in an automated process and returns the data structured in a spreadsheet." Maeztu makes it very clear that "under no circumstances does the patient data leave the hospital's virtual environment. The whole chain of privacy is always maintained."

Once the information is structured, it is possible to predict, for instance, whether a patient will have complications or not when undergoing surgery. "For example, Vall d'Hebron's orthopaedic department uses AI to predict what will happen in a carpal tunnel operation," says Maeztu. "This can be learned from structured data from their medical record but also from the records of other patients who have had the same operation in recent years." With all this information, a prognosis is made which in the end the doctor has to evaluate to decide whether to operate or not.

38 "IOMED i l'Hospital Vall d'Hebron desenvolupen nova tecnologia per a la gestió d'historials clínics" (https://www.europapress.es/catalunya/noticia-iomed-vall-dhebron-desarrollan-nueva-tecnologia-gestion-historiales-clinicos-20180211115250.html).

2.4.2. Justice system

The use of artificial intelligence in the judicial field is a very sensitive issue because just as in healthcare, the decisions taken by the algorithm directly impact people's lives. Similar experiences in the United States, where the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) program decided on criminal reoffending, does not boost confidence in the machine's results either. Independent media outlet *ProPublica* reported[39] that COMPAS, which had been developed by Equivant,[40] featured biases that set a higher probability of committing crimes for black defendants than for white defendants.

However, automated systems have evolved a lot in recent years. And while keeping biases in mind, consideration should also be given to the prejudices of judges (on the basis of race, gender, religion, etc.) and how they may influence decision-making. It would therefore be appropriate to ask ourselves: who will be fairer, a judge or a machine?

Human Decisions and Machine Predictions[41] is an attention-grabbing American study which shows how when deciding whether to give bail in legal proceedings, machine learning can work better than a judge's decisions even when they might discriminate against blacks or Hispanics. The results showed that when it was very clear that the risk of the prisoner reoffending was very low, the judges and the algorithm agreed to release them on bail before the trial. However, the machine was fairer than the judge in predicting cases with a higher risk of criminal reoffending. And this is because machines are systematic, even when they are as racist as judges.[42]

In Catalonia, programmes similar to COMPAS have been used for about ten years to identify criminal reoffending in adults and young people. To date no research has shown that there is any detrimental bias against inmates. Researcher Carlos Castillo,[43] Director of the Web Science and Social Computing research group at Pompeu Fabra University (UPF), has conducted several research projects on these smart systems and in his view they work pretty well: "The specialists who use them ultimately evaluate the results produced by the machine individually and decide on the measure to be taken."

39 "How we analyzed the Compas recidivism algorithm", *ProPublica* (2016) (https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm).
40 Equivant (https://www.equivant.com/).
41 *Human Decisions And Machine Predictions* (https://academic.oup.com/qje/article-abstract/133/1/237/4095198?redirectedFrom=fulltext).
42 Ricardo Baeza-Yates, Karma Peiró."És possible acabar amb els biaixos dels algorismes?".
43 Web Science and Social Computing Research Group (https://www.upf.edu/es/web/etic/entry/-/-/24095/adscripcion/carlos-alberto-alejandro-castillo).

• Predicting criminal reoffending

Release on temporary licence is used for the reintegration and rehabilitation of inmates. Being able to forecast with the greatest possible predictive effectiveness the likelihood of future infringement of a licence is a great help to prison staff. RisCanvi[44] is a protocol (or risk assessment tool) which was implemented in 2009 in all prisons in Catalonia to estimate the chances of a person going back to crime once they had left prison. The Catalan Ministry of Justice commissioned it from the Group of Advanced Studies on Violence[45] (GEAV) at the University of Barcelona (UB) and in the time it has been in operation it has already been used with some 20,000 prisoners.

The crime prediction made by the algorithms is individualised and personalised. "At one time or another in prisons, specialists, governors, psychologists or lawyers have to make the decision about what might happen on an inmate's licence or when they are about to go back out on the streets for good," explains Antonio Andrés Pueyo,[46] senior researcher at the GEAV and Professor of Psychology at the UB. "The concern is whether they will commit a crime again."

Pueyo says that traditionally analysis was carried out based on a few parameters and whether the inmate met certain requirements, and they took a decision on the basis of the results. "This process has been automated for the last ten years with artificial intelligence," adds Pueyo. "Using 43 variables which a mathematical system combines seamlessly, the specialist can take the best decision. If a prisoner asks for a licence to visit their family and RisCanvi's response is that there is a high probability of offending, it notifies the specialist with a red signal. The prisoner leaves anyway, but daily follow-up, an electronic tag or contact with a family member is activated."

The program also tracks the inmate's violent behaviour in prison: whether they have assaulted another person, attempted self-harm or suicide, etc. It is done individually for each inmate and is standard practice in the prison system. However, the specialist no longer has to retain all this information as it is kept by the automated system, which means they have a great deal of input when they have to make a decision.

RisCanvi is now in its version 3.0. Every three years it is updated and improvements are introduced to make it more accurate. In the case of potential ethical dilemmas, Pueyo points out that "the algorithm's response is always validated by the Treatment Board which can: a) maintain the same level of risk (low, high) as the algorithm; b) increase it; and/or c) reduce it. In cases b and c, it has to justify this change with the evidence that supports it."

44 RisCanvi (http://cejfe.gencat.cat/es/recerca/cataleg/crono/2017/eficacia-del-riscanvi-2017/).
45 GEAV (http://www.ub.edu/geav/).
46 Antonio Andrés Pueyo (http://www.ub.edu/personal/docencia/profes99_2000/pueyoficha.htm).

• Algorithms to identify youth reoffending

The Structured Assessment of Violence Risk in Youth (SAVRY) program[47] takes the same approach as RisCanvi but was developed in 2003 in the United States and not by GEAV. "It is used in many countries around the world: Canada, the USA and many others in Europe such as the Netherlands," explains Pueyo. "It has fewer valuation factors, only 26, but it is also used in a highly individualised way. It is more manual than RisCanvi and the final assessment depends more on the specialist. That's because in the case of young people, behaviour changes are very abrupt or quicker than in adults. So that's a good thing. Automating SAVRY much more would not be a good strategy."

When the reoffending risk assessment is high in both RisCanvi and SAVRY, the treatment teams decide on the actions or measures to be taken to prevent what the machine suggests from happening.

• Predictive statistics for lawyers

There are case law tools[48] that also help lawyers in the gruelling task of reading sentences and combining them to draw fresh conclusions. Mathematical models can operate with qualitative information from analysis of millions of cases, from judges who have handed down sentences, from the application of articles and laws, date, place, etc. The algorithms combine all the information from court orders and courts throughout Spain and specify the most appropriate procedural strategy for each case.

• Guidance on the extradition or otherwise of migrants

When an illegal migrant is apprehended and brought before a judge, there are certain criteria by which the judge can decide not to return them to their country: risk of death, risk of getting a serious illness, an epidemic, or the risk of being subjected to torture or cruel, inhuman or degrading punishment.[49] The judge gives each of these criteria a score and then makes a decision.

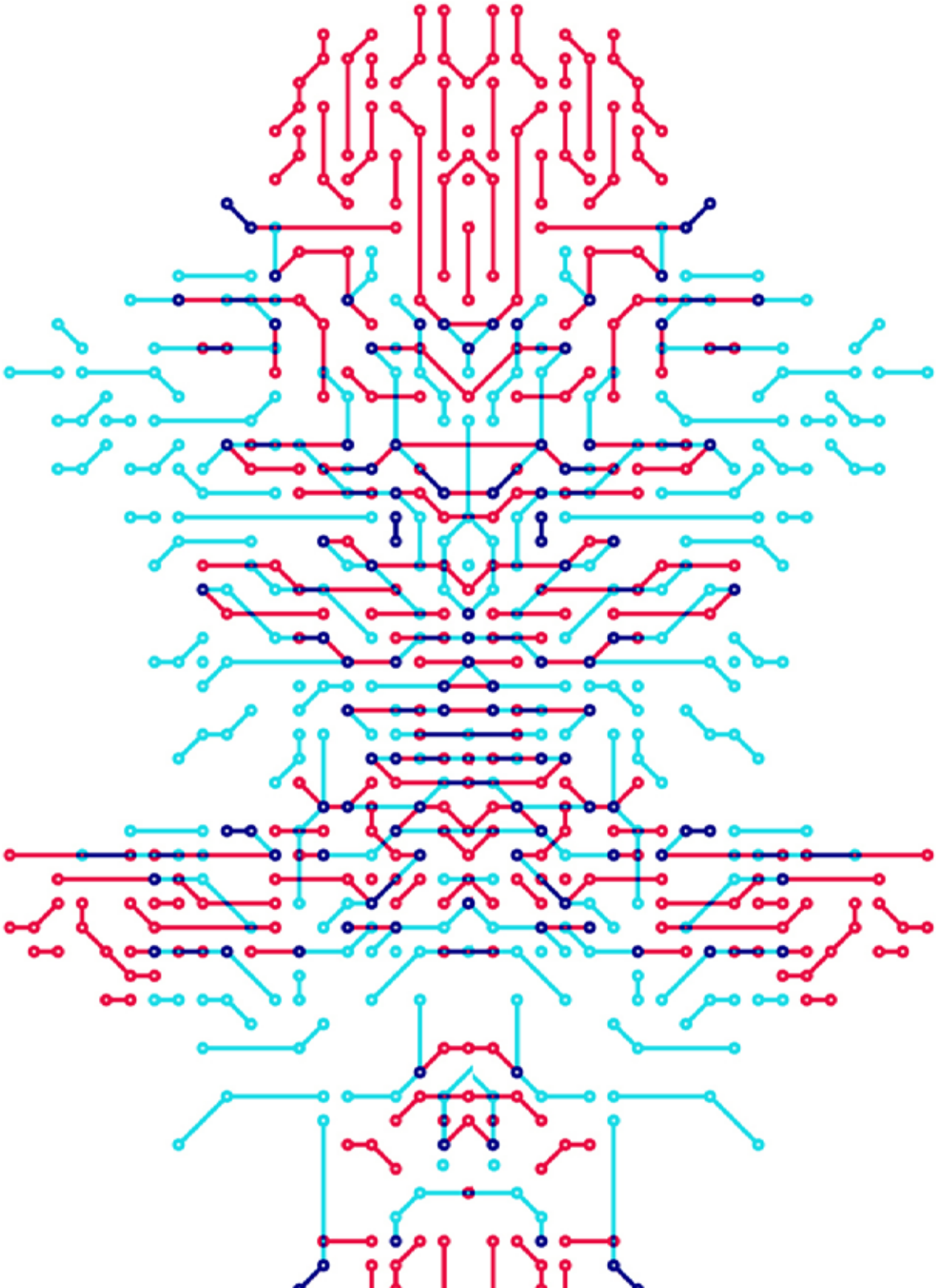
The UB's Department of Mathematics and Computer Science drew up a project with all the cases that had gone Spain's Supreme Court and created a mathematical model. Jordi Vitrià, a researcher at the UB, explains that the tool gives the lawyer very valuable information for associations that help migrants. "The data is all from real cases and the

47 SAVRY (https://treballiaferssocials.gencat.cat/web/.content/03ambits_tematics/07infanciaiadolescencia/temes_relacionats/jornades_treball_dgaia_2012/docs_3_maig/valoracio_risc_reincidencia.pdf).

48 Jurimetria (https://jurimetria.wolterskluwer.es/content/Inicio.aspx).

49 *Migration, HR and Governance* (https://www.ohchr.org/Documents/Publications/MigrationHR_and_Governance_HR_PUB_15_3_SP.pdf).

tool provides guidance on how the extradition might go," Vitrià adds. "It is guidance for the lawyer and could also be for the judge, because the model is created based on the Supreme Court cases. But it is evident that the final decision is up to the judge." As yet it has not been implemented in Catalonia.



2.4.3. Education

Artificial intelligence has revolutionised all industries and education has not been left behind. Last May, UNESCO brought together 50 ministers in Beijing[50] to agree on the document *Artificial Intelligence in Education. Challenges and Opportunities for Sustainable Development*.[51] One of its main points is the message that AI has the potential to profoundly transform education while always respecting human rights and social values. Education policies should also be planned with this approach in mind so that young people can take ownership of their future. Studying only when you are young is over. Learning is now something that will last a lifetime.

The experts say that AI will greatly help the shift in education from task-based learning to collaborative learning. Moreover, the processing and analysis which can be done today for assessments, distance learning platforms, interaction in class with mobile phones and browsing through educational websites allows us to envisage new study patterns. Algorithms can now establish relationships and adapt to each student by predicting the best way for them to acquire knowledge.

In primary and secondary education classrooms the automated system can guide the teacher, while in universities it can manage research and make faculty resources more efficient.

• Active learning

The Assessment and Learning in Knowledge Spaces (ALEKS)[52] programme has been tested in the United States for over 12 years with millions of students and now is reaching Catalan schools. “We should not be afraid to introduce AI into the classroom,” says Carles Sierra, Director of the Institute for Research in Artificial Intelligence (IIIA). “It can be a tool that improves the learning process.”

ALEKS uses adaptive questioning to quickly and accurately determine exactly what a student knows and doesn’t know in a course. It also advises the student on the topics they are most ready to learn. As the course progresses, ALEKS periodically reassesses the student. It can be used to plan personalised learning for each student or for ones who have greater difficulties in learning a particular subject.

50 International Conference on Artificial Intelligence and Education (https://en.unesco.org/events/international-conference-artificial-intelligence-and-education).

51 *Artificial Intelligence in Education. Challenges and Opportunities for Sustanaible Development.* UNESCO (https://unesdoc.unesco.org/ark:/48223/pf0000366994).

52 ALEKS (https://www.aleks.com/).

"The system makes a model of each student to map the details of their knowledge," adds Sierra. The system provides the teacher with a pie chart to track the student's acquisition of concepts. "We should always bear in mind that it is a support for the teacher in the classroom," he notes. "Evaluation studies have shown that university dropout rates are lower and marks at secondary school improve with this technology."

• The algorithm makes the class groups

In a 30-strong class of students who have to carry out a team assignment, an automated system groups the students by their personalities. "The teacher sets the number of students per team and the skills required for the task," explains Sierra. "The AI groups and assigns a responsibility to each student."

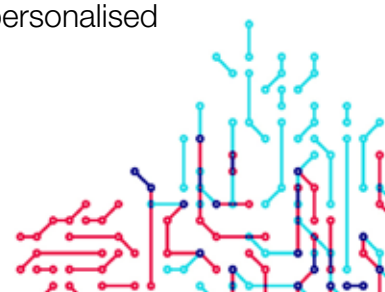
But how can a machine set up the teams to carry out an assignment without making a mistake in the groups? "First the personalities are evaluated using 20 questions and the smart system classifies them: introverted-extroverted; sensitive-intuitive; analytical-emotional; decisive-reflective," says Sierra. "Depending on the result, the automated system will group them in the most diverse and efficient way to tackle the exercise provided by the teacher." The result of the algorithm may or may not confirm the teacher's opinion of the students. Sierra says the idea is to find balanced teams and ensure all of them are good enough: "This is the difficult part. Finding a good team is easy, but making sure they are all at least up to an acceptable standard is very hard. And that's what the algorithm does."

Several high schools in Catalonia have already tested this technology and Sierra says performance improved by between 25% and 30%. Ethical concerns might arise from these applications. Is there no risk of the machine making a mistake and a student being harmed? "The results we have now show that on average it works well," he replies. "This does not mean that there are no errors in any particular case. But we have to bear in mind that teachers are not free of bias either and they can also make mistakes when dividing students up into groups."

• Marking exams

Another technique the Institute for Research in Artificial Intelligence has put into practice with high school students is what is known as *peer assessment*.

A tool has a pre-set scoring guide for how students should be assessed. The teacher only assesses a few students but eventually the whole class has its own personalised mark. How is this possible?



"Each student has a group of classmates to assess. What do we do? We compare the way the students assess and that yields a degree of similarity in the assessment," explains Sierra.[53] How can a student have enough judgement to assess a classmate? "The smart system creates similarities between the teacher's way of assessing and the students' way of assessing. The algorithms build a level of confidence between the teacher's assessments and the ones made by the assessing students."

This experiment was done in an English class and the results were no more than 10% off the marks the teacher would have given.

• Tackling dyslexia

For a couple of years now, some 70 schools in Catalonia have had an artificial intelligence system to help students with dyslexia or reading difficulties. In total, about 3,000 children aged 5-12 benefit from it along with another 200 who are closely monitored privately.

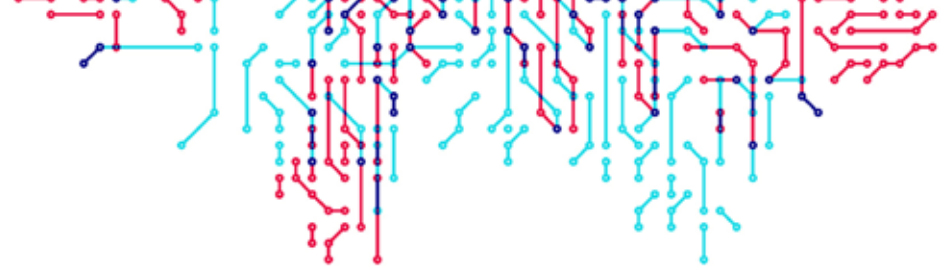
UBinding[54] is a learning platform based on the cognitive development of each child, the family and the school setting. A group of University of Barcelona (UB) researchers developed it in 2007. Its promoters say the UBinding method helps 90% of students to improve reading fluency in 7-8 months on average.

"This is a decision support algorithm," explains Jorge López, head of the research and development unit. "A team made up of psychologists, speech therapists and mathematicians specialised in learning disorders perform monitoring on a daily basis." It could also be used with children with attention deficits because they fail in executive functions or to improve remedial work in some subjects.

"Training allows us to adapt to each child in line with the speed at which they assimilate reading," adds mathematician Adina Nedelea. "As there are thousands of students, it is very difficult to provide individualised monitoring without the help of artificial intelligence. The algorithm never decides but rather suggests what the student's next session should be like and the professional decides whether or not to do it."

The UBinding team gets feedback from the parents and the child who has done the reading exercises. "Without the algorithms we would not have been able to reach thousands of students," points out Nedelea.

53 Carles Sierra. "Intel·ligència artificial i educació" (video) (https://www.youtube.com/watch?v=Xd5ZoKdl33A).
54 UBinding (http://www.ubinding.cat/).



• Virtual assistants in the classroom

The most cutting-edge machine and deep learning programs can create a personalised study method for each student based on the data generated by their participation in the classroom or at a distance. The algorithm recommends the study pathway according to each student's pace, skills and objectives. It can also suggest the best classmates to do a project with by analysing the compatibilities and skills of each group member. "In a few years, the classroom will be completely different: the teacher will have a virtual assistant to help them with what they have to do, the students will have a personalised programme and class attendance will be face-to-face or virtual," claims Ivan Ostrowicz,[55] an expert in artificial intelligence applied to education.

All the students learning with the same textbook is already old hat. Ostrowicz says the algorithms will recommend content for each student consistent with their level while dashboards will help the teacher to check each student's progress and virtual assistants will enable students to get answers to their questions whenever they want. Likewise, they will have a recommender of the best students to learn with and subjects tailored to each of them. Yet the most exceptional of all these promises is that an algorithm will be able to warn the student when they are about to forget what they have learned. The adaptive retention system measures the speed of forgetting and recommends a review of the lesson just before what has been learned is lost.

• Facial recognition for entering a high school[56]

This example is not an application of AI to improve learning but rather a facial recognition system which tracks the presence of students in a school.

It had been running for seven years in a high school in Badalona and most families were happy with it. "The first year was a shock, but it's going very well because if your kid doesn't go through the facial recognition system, they text the parents," said one of the mothers in a TV3 interview. Using cameras mounted in the facility the students had to "clock in" when they got to school; otherwise, the parents were notified of the child's absence. The system did not monitor class attendance, although if the student was in the school they would surely go to class.

Professor of Law and Political Science Mónica Vilasau Solana[57] says biometric data and data on minors are flagged as particularly sensitive. Consequently the Catalan Data

55 Ivan Ostrowicz (https://www.linkedin.com/in/ivanostrowicz/?locale=fr_FR).

56 News item posted on 324.cat (5/10/19) (https://www.ccma.cat/324/reconeixement-facial-per-passar-llista-en-un-institut-de-badalona/noticia/2952712/).

57 Mònica Vilasau (https://www.uoc.edu/portal/es/news/kit-premsa/guia-experts/directori/monica-vilasau.html).

Protection Authority launched an investigation into the issue. “It is not enough to have parental consent; the impact of using them has to be evaluated and we need to find out whether there is no other alternative available,” says Vilasau.

Other Catalan schools had implemented facial recognition of students, but following the European General Data Protection Regulation they had to disable the system. The high school in Badalona also ended up removing it as a result of the investigation begun by the Catalan Data Protection Authority. This example clashes with the most ethical aspect of AI: in these cases the technology was used to control and not to improve performance in the education of young people.

2.4.4. Banking

Banks have always been among the first to apply automated decision-making algorithms as far back as the 1970s, albeit for much simpler transactions than the ones today. Now they are advanced neural networks which offer an enormous variety of financial products depending on the data entered: common options include whether to promote a service through a customer's email inbox and recommending products when people visit the financial institution's website. Artificial intelligence is also used for internal tasks in banks to gain efficiency and performance, handle customer calls, achieve acceptable phone waiting times, design priority service for premium customers and decide where to site new branches in a town or city.

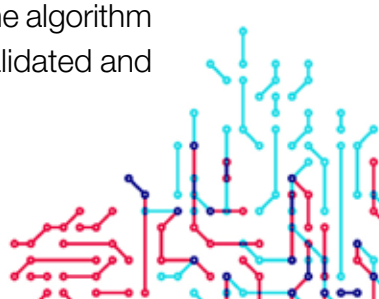
Banks are strictly controlled in terms of the privacy and anonymisation of the data that are pumped into their algorithms. So right from the outset the principle of not harming any customer with potential discrimination generated by biases prevails. "It also has to be remembered that a bank makes its money by discriminating, although the word has a negative connotation," explains Marco Bressan, a former lead data scientist at a nationwide bank. "The institution needs to know whether it will get back the money from a mortgage or a loan. Using algorithms to decide who will or will not be given one is essential. Meanwhile, the bank has to identify and mitigate any biases (gender, income level, customer origin, etc.) there may be."

Bressan is convinced that algorithms when well designed are more accurate than people. "Let's picture a misogynist bank manager: he'll probably make some very bad decisions because of the personal prejudices he has. Automated systems react to a number of variables and the decision they make is fairer," he says. "The data the algorithms are given is the most important thing. If there is initial discrimination, they will perpetuate it over time and it will impact many more people."

Furthermore, in Bressan's view humans can no longer calculate the amount of transactions in real time that machines can do with big data. "However, as they operate on a large scale, the algorithm only needs to fail once to affect thousands or millions of people at a time. By contrast, when the bank manager was wrong it only harmed at most a hundred people."

• Loans for whoever has the best score

The most heavily regulated area of a financial institution is access to credit. The algorithm that initially decides which type of customers are given or denied a loan is validated and



regulated by each country's central bank. However, the customer is not aware of this assessment.

How does an algorithm decide whether or not to give a loan? When a bank lends money, it makes sure it will get it back. In the past, information was gathered from the customer and the bank manager took the decision. With more powerful technology, the forecast of the cash which will be recovered or lost is more accurate. The bank knows how much money the person earns and how much they can save each month after paying for their outgoings. Customers are classified into clusters and assigned a score that is not necessarily related to what they earn. A person on a salary of about €1,500 per month can save €100 and get a higher score than someone who earns €4,500 and is not able to save even €50.

How are these algorithms trained to be effective? Let's say we have the data of thousands of customers who have been given loans in the past. Based on a number of variables, the algorithm is programmed to predict whether or not these people are sufficiently creditworthy to repay them. As these data are drawn from the past, they are easily comparable in order to find out whether the automated system is working properly. It has already been shown that accuracy is always higher with an automated system than when done by hand.

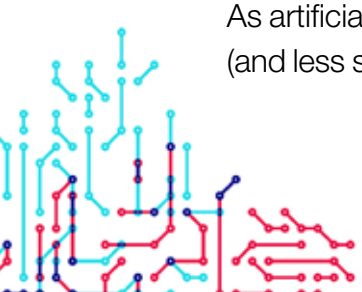
• Taking out insurance

Scoring systems are common in insurance (vehicle, health, home, business, etc.). A young person is obviously more likely to have a traffic accident than a 55-year-old and therefore the cost of their insurance will also be higher.

We might ask whether these automated systems are never wrong: the answer is that error or accuracy will depend on the previous information (big data) that is entered in the algorithm. When predictions about groups are wrong, more people are harmed. It may be the case that a young man is more cautious than an older man and that young people always end up paying more. Insurers' algorithms factor in other variables when deciding the premium such as income. It has been shown that people on lower incomes normally make more claims than those with higher earnings. Hilly terrain is also factored into the price they have to pay because it greatly influences accident rates.

• Mortgage predictors

As artificial intelligence becomes more powerful, algorithms become more mathematical (and less statistical). "This has led us to shift from descriptive to predictive analytics," says



Pier Paolo Rossi, Director of Advanced Customer Marketing & Analytics at a Catalan bank. “We study the past to predict the future. The algorithm analyses, for example, all the customers who have taken out a mortgage in the last five years and based on some variables it predicts who might potentially take out a mortgage in the immediate future.”

The variables may be socio-demographic (age, gender, children, etc.), financial products taken out (current account, mortgages, loans, etc.) or how they interact with the bank (whether or not they have a credit card, whether they use it a lot or a little, whether they always take money out of an ATM or go to their branch, etc.). The relationship with banks in small towns and villages is closer than in cities and this is another variable. The data are used to train the algorithms which will make the predictions with a very high probability of getting it right.

Prescriptive analytics is also starting to be brought in. Algorithms identify customer needs based on their lifecycle. A couple who are 25 years old, living in a rented flat and have no children are different from the same couple ten years later. If everything has gone okay for them, they will probably have a higher salary, they will want to buy a flat, they will have children and their needs will be different. “With this analysis both the bank and the customer win because whereas before mortgages were offered to everyone, now they are only given to people who need them,” says Rossi. The data do not slip up.

• Financial recommenders

As we have seen in the examples above, banks have an advantage over other companies because they have a lot of information about their customers. Data are money. To find out which financial products will interest them, banks only have to conduct *descriptive analytics*: in other words, examine what their customers are like, how many make it to the end of the month, how many are from the wealthier part of town, how old they are, what their lifestyle is like, etc. The algorithms help customers to make financial decisions, buy on credit, manage household or family outgoings, take out insurance, etc.

The banks’ business has changed considerably in recent years and is no longer just about lending and holding money but rather about offering services that set them apart. How? By using information about the life habits of thousands of people including the area of town where they live, their type of job, salary, family members, children in their care, etc. And also by using spending data: on electricity, water, gas or any other household outgoing, daily shopping, travel expenses, studies, extracurricular activities, holidays and plane and hotel bookings for work or leisure. Drawing on all these data the algorithms paint an all-inclusive picture of each person’s life and offer them the services that *may be of interest to them*. By doing this they enhance each customer’s loyalty.

• Financial *coaching*

As spending data are the most valuable, they are used to conduct anonymised performance analysis which classifies the customer into a certain group. This makes it possible to identify products which all of them would buy. In the previous example we have already seen the income bracket and monthly outgoings variables which can be factored in.

Banks send and/or make extremely personalised purchase offers which the customer is unlikely to reject because they will save money on a service they already use or which will soon become an advantage for their family. In other words, if a person the same age as me, on the same salary and with the same outgoings has Wi-Fi at home which is much cheaper than mine, the bank will suggest I switch providers and take out a new service through the financial institution. These personalised offers are aimed at hundreds or thousands of customers.

The automated financial recommender replaces the traditional bank manager, the person who was trusted to open a pension plan or make long-term savings. Let's say a person has €10,000 they want to save for their retirement or invest to get a return. The algorithm will classify them according to whether they have an "aggressive" risk profile (they are prepared to accept high risks) or whether they are rather "conservative" (low risk) and will then tailor the bank's proposal to their preferences. These types of assistants without intervention by a human help both the bank's staff and the customer: they warn the customer if there is a risk they will lose their money, and even if the bank does not gain anything from the transaction it nevertheless builds up loyalty.

These are measures designed to restore confidence in the banks which was severely shaken by the financial crisis. Another operation that builds loyalty is to tell customers when they will soon be in the red. As the automatic system monitors the accounts, it can warn of an awkward situation that everyone would like to avoid. And it recommends curbing outgoings on restaurants, leisure, non-essential purchases, etc.

• Credit card fraud

Has your credit card ever been rejected in a store where you had never shopped before? Or has it been blocked because you have spent more than usual? Although we are not aware of it, this is more common than we might imagine: every day millions of cards are blocked all over the world.

Machine learning has been used to detect financial fraud since the 1990s and is now very sophisticated, yet it is still not completely perfect. The algorithms are trained to block credit cards when they identify odd or inflated spending patterns. And they do this by monitoring millions of transactions every day.

But what if the automated system makes a mistake in viewing a transaction as odd when it isn't? This is known as a *false positive*; i.e., a mistake in the prediction and in the decision to block the card. A study conducted in 2015 by consultants Javelin Strategy Research[58] estimated that only one in five fraud predictions is correct and that errors by the bank could cost it billions in lost revenue as people who have had their card blocked do not use it again. Although automated fraud detection is very sophisticated, it does have some limitations since in order to detect real cases of fraud the algorithms have to get it wrong very often.

To avoid further frustration for customers and headaches for banks, researchers at the Massachusetts Institute of Technology (MIT) used a technique,[59] which is already employed by some banks in Catalonia, to almost halve the possibility of error. This system looks at other characteristics of the purchase which had not been factored in up until now such as the distance between two shops, the time the purchases were made and whether both were made in person or online.

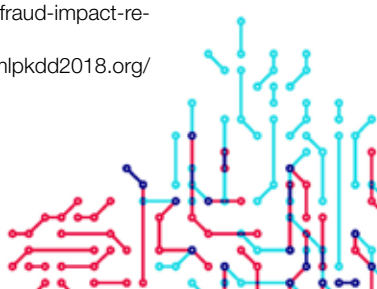
• Algorithms that interact with the customer

Online virtual assistants or *chatbots* are algorithms which interact directly with customers, simulating a bank manager. The written conversation is recorded as a chat along with their visits to the options available on the website. Cookies allow the bank to learn where a person has been during the time they have been on the website.

"Based on what the customer has looked at, the questions they have asked the chatbot and on the answers it has given, the algorithm learns and can personalise the customer's screen then and there for what they need," says Rossi. "This means we know whether they are interested in a loan, whether or not they are a customer, how much they need and what for (a trip, a property, education, a car), and the algorithm personalises the answer on the spot." This system is also used by Amazon or any online store. "If you know your customer, you have an advantage over others," he adds. "And you generate trust you didn't have before."

58 *2015 Data Breach Fraud Impact Report* (https://www.javelinstrategy.com/coverage-area/2015-data-breach-fraud-impact-report).

59 *Solving the false positives problem in fraud prediction using automated feature engineering* (http://www.ecmlpkdd2018.org/wp-content/uploads/2018/09/567.pdf).



2.4.5. Business

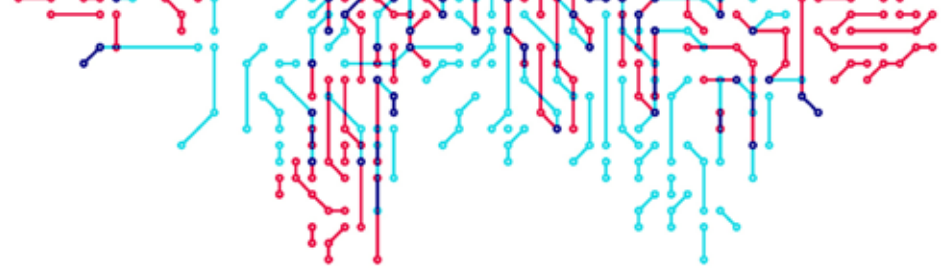
Without our noticing it, we are surrounded by algorithms which suggest products, recommend a series, rent us a flat, offer us a second-hand car, etc. And they always get our preferences right.

Similarly, we have allowed small devices into our lives which greet us first thing in the morning or with which we interact to ask for music, turn on the TV while we are cooking or to tell us the weather forecast. They are the virtual assistants, micro-intelligences which understand natural speech. We also find them as chatbots, often on websites, and they interpret what a customer or a member of the public asks them. Based on the information they receive, they choose an answer. These systems save time for a company's customer service department, but they also replace workers who in the not so distant past took calls and answered queries by phone or email. But what we have today is very early artificial intelligence. We have not yet seen anything of what is to come.

In Catalonia, tech multinationals are taking over the market and being competitive in artificial intelligence is not easy says Jordi Navarro,[60] CEO of a company engaged in predictive analytics and machine learning for the last four years. He believes that the tourism sector still has some way to go. "Using historical information, we can understand the past and extract patterns. This allows us to operate in the hotel sector, in hotel room cancellations, for example. With AI, in a way it's as if we had a crystal ball which enables us to operate with a fairly high degree of accuracy. Human behaviour is very predictable. Otherwise, why is the world leader in ecommerce so successful?" he asks.

Jordi Mas is a member of Softcatalà and a pioneer of the Catalan Internet. He is worried about the ethical side of the deep learning systems used in business settings, especially because how a machine makes certain decisions cannot be explained. "I am concerned that there is no code of ethics observed by everyone. For example, no professional should work on a business project that does not respect international human rights conventions. I am also concerned that regulation is lagging behind technology and that we are too reactive. As is the case now with facial recognition, that we are not aware that biometric data is extremely valuable. The algorithms choose the content we read, watch or listen to, the products we end up buying, filter out what we are not interested in, and a thousand other things. There is a danger of creating simplistic bubbles in a world that is complex and full of options. I am concerned about unethical companies,

60 Jordi Navarro (https://www.linkedin.com/in/jordi-navarro-perez/).



yet also very much concerned about governments. You only have to remember that much of what we have comes from the United States or China."

• Room cancellations

In 2017, a United Airlines passenger[61] grabbed all the attention when he was forcibly ejected from an aeroplane. There had been overbooking (the sale of tickets exceeded the aircraft's capacity) and someone had to stay on the ground. That person was the customer who had built up the fewest air miles (or flights). Another passenger taped the incident and the video went around the world. The company reimbursed the tickets of everyone who was travelling on the unfortunate flight 3411.

This situation is not new. Overbooking has always been around. It used to be done by eye, by intuition or industry knowledge, and now it's done with mathematics and AI. Navarro explains that the same strategy is used in the hotel industry for room reservations and cancellations. "Now we make fewer mistakes," he says. "Everything is based on statistics using the profile of the person who has made the booking, their behaviour on previous occasions and other variables. The algorithm is able to predict whether or not they will cancel their booking at the last minute. The hotel owner can offer rooms but doesn't lose money."

• A click doesn't mean a customer

Often when we are looking for accommodation in a tourist destination, we view the hotels on an intermediary price portal or website. We enter the town and the number of nights in the search engine and it gives us a result. These intermediary websites are wholesalers that offer everything the establishments have, just like the travel agencies of yesteryear. What is the wholesaler's business? The hotels' advertisements. For example, if someone is looking to spend two nights on Menorca, the wholesale website will offer a list of accommodation options with their different prices. Just clicking any of the offers means the owner of the establishment has to pay the wholesaler. If the room is not booked later on, they have lost money. And this happens thousands of times every day. The hotelier therefore has to be convinced that what they are offering to the wholesaler is attractive enough for the clicks to result in room bookings.

Now with AI you can be much more accurate depending on the town, the time, whether they ask for one or two beds, the dates, etc., and assign a high booking likelihood. Thus

61 "United Airlines reembolsará los billetes a todos los clientes del vuelo del que expulsó un pasajero". *20 Minutos* (https://www.20minutos.es/noticia/3011324/0/united-airlines-devolvera-billetes-pasajeros/).

the hotelier only offers the wholesaler a room in Menorca when they are almost certain that the customer will keep it. Could this be done without algorithms? “No,” replies Jordi Vitrià, a researcher at the UB’s Department of Mathematics and Computer Science which has collaborated on this project. “It’s a question of scale. In the past, the traditional travel agency had contacts with hotels, car rental companies, airlines, etc., and it kept a margin for everything it offered and sold. This would be the role of the wholesaler. Now, in real time, 24 hours a day, it is impossible to do it personally.”

Is it ethical that if you log in from the wealthiest area of the city, a room will cost you twice as much? “Business is business,” says Vitrià. “Even though you aren’t playing with clear rules because nobody has told you that they have taken your geolocation data and that this may influence the final price you pay.”

• More affordable MOTs

In France, how much you pay to pass your vehicle’s MOT is deregulated. As there is a lot of competition between the franchises that inspect the cars, the University of Barcelona partnered a project to set automatic prices based on where the inspection facility is. “If there are seven competitor MOT facilities the situation is very different from you being the only one in the area”, explains Vitrià. “We built an algorithm that used all the data from vehicle inspections in France in recent years to automatically suggest to the MOT franchise manager the price they could make competitive depending on local circumstances or the day of the month, etc.”

• External factors for getting production right

Large vehicle manufacturers need to have very tight control over their supply procurement. Some parts have to be ordered well in advance as they are sourced from Asian countries. The carmaker’s head office makes forecasts but they often do not match up to reality because they do not consider other non-production factors. This leads to high inventory in warehouses which in turn means money tied up.

Production predictions can be improved with a mathematical model says Vitrià. “The results are very different using supply history and also the weather forecast, bank and other holidays where each warehouse is and data about socioeconomic factors. This is information which is never factored in when manufacturing cars but it is extremely influential.”

• How to get the most out of weddings

The customer journey is about tracking the customer from the moment they contact a company (to ask for information, for example) until they leave it (deregister from a service). Based on the interactions you can predict when that customer will be lost. If you sign up for a scooter hire service for getting around town, for instance, and do not rent one for a while, the company will email and text you with deals, savings, opportunities and routes to take until you rent a scooter again. And if you don't react, they will consider you lost.

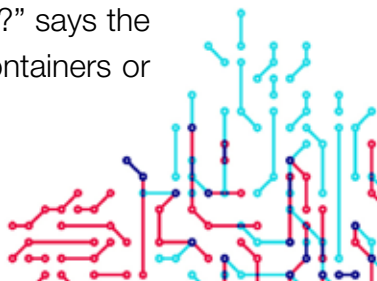
Most companies operate with this strategy today. In Sant Cugat there is one that offers all kinds of services and products for weddings. It is the huge wedding search engine: from makeup to restaurants, special car rentals, dresses, etc. Over a hundred people work there to answer questions and queries in any language. It is in Catalonia but its market is global.

The business works just like the hotel room reservation business. The company earns money every time people click a service and the advertising restaurant pays it. The restaurant shells out to be listed first in its search engine. If you come out on top, there's a good chance people will make wedding bookings. "Although births are an important time for most people, they don't trigger as much expectation or shift as much money as weddings," says Vitrià. "The algorithms help to enhance this company's customer journey and tell it when there is a risk that the customer might leave, when something should be done in person or by email."

• Algorithms to be faster

Bike and motorcycle delivery companies also use algorithms. Small and very diverse things are asked for in last mile delivery. These companies rely on the algorithms for efficient order and delivery management. For example, they assign the delivery person who will meet the customer's order first and tell them where to buy what the customer wants based on the distance to be covered and the time set. If the distance is overly long, the application will transfer the order to a delivery person on a motorbike. This is also a benefit for the delivery person as it means the customer doesn't have to wait as long and may give them a better rating at the end of the service.

The algorithms assign the delivery people a score based on their reputation using criteria such as time spent on the delivery, whether it is at the weekend, whether it is at peak demand time, whether they have no customer complaints, etc. With machine learning, the system can also understand what customers are asking for and how they are asking for it. "For example, if they ask for 'three waters', what does that mean?" says the communications manager at a home delivery company. "Three five-litre containers or



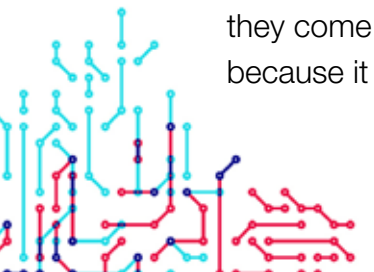
three one-litre bottles? The platform has a database with all the accumulated information and processes it so that delivery is as precise as possible."

• Behavioural data; the most valuable

We all visit portals featuring ads for buying or renting a property, second-hand cars, job offers or to watch a series. But how do the algorithms behind them work? The main objective is to have maximum user engagement on the platform or the website. This is the business: the more time spent on these portals, the more chance they have to make money from the ads, products or services they offer. With machine learning, recommendations are made that are almost always right, fraudulent ads (and there are many) can be identified and they can interact with the user by giving smart replies.

"Recommendation algorithms were invented twenty years ago so they're nothing new. They learn from user behaviour," explains the head of an online marketplace, an international company based in Barcelona and owner of several advertising portals. "However, we have two problems: 1) Introducing new products or services. If you have a portfolio featuring 7,000 products which already work on their own, but there's one that nobody has heard of because it's new, how do you get people to look at it? You train the algorithm to recommend it until it grabs the user's attention. 2) What in the industry we call a 'cold start'. We don't know anything about this new user, we have no behavioural data. Maybe they have browsed twenty pages, but they haven't bought anything. The algorithm does not know what they like, so it is in the dark in its recommendations. If it gets them to buy something or put a product in their basket, that's already a lot of information! A person usually spends twenty minutes on the site, which shows them very generic things."

All ecommerce companies operate in the same way. When a person gets to an ad portal website, they are sent a cookie (a small file that identifies them with a unique number). Cookies last over time and they provide a lot of information for months or years. What data are gathered? The categories visited, searches, pages viewed, time spent on each page, purchases, etc. "If I have a user who goes to the book section every time they enter the portal, conducts searches in Java, Python, HTML, has visited me a thousand times but has not registered and has not bought anything, I still have a lot of data!" adds the marketplace manager. "We analyse all this information and every time a user is identified, the recommenders are triggered. We make millions of recommendations every day for all the portals we have. They say things like: 'If you've seen this bike, you're bound to be interested in this helmet.' Anonymous ones count too. Maybe they come in today and don't come back for six months, but they still have the cookie because it doesn't expire."



2.4.6. Social

Artificial intelligence in the social sector is very useful, especially for large-scale processes. Algorithms can award or deny financial benefits, assistance of any kind and refer to other specialised services depending on the personal problem. Nowadays they are already very helpful, particularly in large cities when thousands of members of the public ask for services.

As noted above, algorithms are fed or trained with big data from the past. And these data are very likely to have biases because societies evolve and habits, customs, ways of living, getting an education and finding a job have changed. Catalan society today is not the same as it was 50 years ago.

For example, women did not used to go out to work as much, there were practically no single-parent families, there was not as much adoption as there is today, etc. We may have similar problems (economic, immigration, violence, etc.), but the realities are different. So when applying AI in the social sector, it is crucial to identify and mitigate the potential discrimination in the big data used and the sooner the better.

In Catalonia, the public authorities are just beginning to get their act together in terms of using AI to redistribute social benefits, but for the time being most of all it is seen as a way of gaining efficiency in procedures and managing resources better.

• Collective intelligence for social benefits

Barcelona City Council's Social Rights Department can handle an average of 50,000 first visits a year. The people who come to the 40 social service centres spread across the city have financial, dependency care, mental illness or alcoholism problems, they may need psychological or adaptation help, they may be experiencing gender violence, etc. These are very diverse issues which are dealt with by a staff of more than 700 professionals, including social workers, psychologists and social educators.

When the person arrives at the centre, they are seen in a private booth. The social worker records the conversation and at the end transcribes the problem along with the help or service to which they have been referred. In the internal system this is described with three letters: need (N), problem (P) and resource (R). Currently the City Council has hundreds of thousands of interviews, many of which end up being repetitive as the problems are similar.

"We went into a repository of 300,000 interviews and we equipped it with machine learning techniques," explains Lluís Torrens, Director of Social Innovation in Barcelona City Council's Social Rights, Global Justice, Feminism and LGTBI Department. "The machine reads all the comments noted by the social workers for the N, P and R. Now it suggests the resources based on what it has learned. It classifies the needs and potential responses."

The results have already been applied in three centres. As Torrens explains, the recommendations are more precise than the ones made by the professionals because they avoid dispersion. And the level of satisfaction is high. "If you have 700 professionals, it's very likely that not all of them allocate resources in the same way," he says. "That will be because their lecturer at university taught the subjects in their own particular manner or because the professional has done more research on a kind of problem, etc. The machine makes the answers uniform and leaves the decision up to the professional. I like to say that it is a collective and not an artificial intelligence system."

• Better evaluations

It is estimated that some 25,000 families in Barcelona live in chronic poverty. Strange as it may seem, no evaluation of the social assistance provided by the City Council has ever been carried out. It is not known whether or not the people or families who have been awarded a benefit or a recovery service have found it useful. "When a doctor gives you a pill, they know what percentage of your illness it will cure. Social professionals do not know whether the action they take will work or not because a lot of factors influence people," says Torrens. "To begin with, if the people who have asked for the assistance no longer come along to the service, it may be that their situation has been sorted out, but also that they have given up on our help or left town."

"AI allows us to make accurate evaluations and compare what the professional would have decided with what the machine decides," he continues. "Now we want to set up an integrated big data system which consists of getting all the information we can from administrative records about a family asking for financial benefits, to find out what their income is from the Tax Office, whether they are getting a pension, whether they have applied for other benefits from other tiers of government, to get a time pattern and find out whether their situation is getting better or not."

The evaluations have been conducted as part of the B-Mincome European project[62] in which 850 families from the city's ten Besós districts have been comprehensively monitored over two years. "We have given them an income with active reintegration policies,"

62 B-Mincome project (http://ajuntament.barcelona.cat/bmincome/ca/pressupost-ajudes-barcelona).

says Torrens. “It’s a very controlled project backed by the public agency Ivàlua. This project is now coming to an end and has given us a lot of insight into how an extensively evaluated action is impacting the families. Public interventions have to be evaluated and the algorithms help us to do that.”

• Identifying bias

After kicking off some pilot tests with artificial intelligence, Torrens’s main concern as Director of Social Innovation in Barcelona City Council’s Social Rights Department is that the algorithms which it operates with should have the least possible bias. So they have commissioned an audit from an external company. “It is very likely that the origin of the families or their gender or age means we unintentionally get results we shouldn’t,” he explains. “And that there is some discrimination due to the algorithm itself, that it hasn’t generated any without our being aware of it. This is to be expected since social responses have changed over time. But we have to keep an eye on biases and mitigate them.”

“We would also like to use the algorithm to identify a problem the social worker may not have picked up on,” says Torrens. “That might be violence in the home (be it towards the wife, the children or the elderly). We would do this by correlating information from other social services, for example the services that support women facing problems of male violence. By joining up the services we could set up an early warning programme and identify a challenging domestic situation. We are still working on this and will explore it further. But it would mean the social care professional could work in coordination with the Catalan Regional Police, the Barcelona Police and other services to help families.”

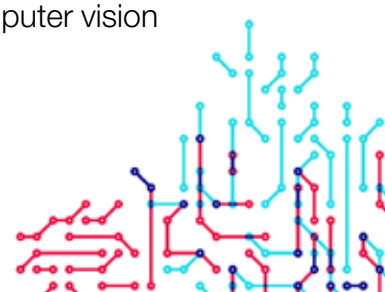
• Studying active ageing

Petia Radeva, the Director of the Machine Learning and Computer Vision consolidated research group at the University of Barcelona (UB), leads the Lifelogging project[63] which consists of taking pictures of a person throughout their life with a handheld camera.

It was funded by TV3’s La Marató telethon and is being conducted together with the team headed by Dr Maite Garolera at the Terrassa Health Consortium. The intention is to monitor daily life in order to study the active ageing of the elderly. Low temporal resolution (LTR) cameras are excellent for this purpose.

As they eat, move, sleep, shop, cook, clean, socialise, etc. older people are associated with cognitive impairment. And it is known that those who keep skills high by performing different activities every day are less physically and cognitively frail. Computer vision

63 Lifelogging (http://www.ub.edu/cvub/egocentric-vision/).



techniques and deep learning are used to analyse these actions based on the images captured with LTR cameras. They are also a scrapbook to improve memory.

• Smart personalised diet assistant

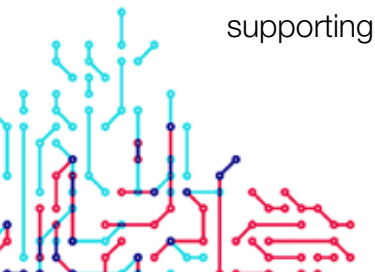
The Technical University of Catalonia (UPC) is taking part in the Diet for You project along with the Hospital del Mar Institute for Medical Research (IMIM) to promote healthy lifestyles and adherence to diet. “The algorithm takes all the patient’s data about their health, lifestyle, their past or present diet, the exercise they do, and in the future their genomic information, etc., and profiles the person,” says Karina Gibert, principal investigator of the project and a member of the Intelligent Data Science and Artificial Intelligent Research Center (IDEAI-UPC). “This means the system has the knowledge it needs to learn about their diet patterns.”

It automatically creates the menus for three or five months ahead with all the dishes prepared which dovetail with the most standard nutritional recommendations. “With the added bonus that the system is configured with what the person likes and dislikes, which ingredients they can’t get because they’re too expensive, what the lifestyle is like in the place where they live (whether lunchtime is longer or shorter, whether people tuck into a quick sandwich or have three big meals, whether they drink tea or coffee, etc.) and information about any allergy or food restrictions such as whether the person is diabetic,” says Gibert. This knowledge is the foundation on which the nutritionally-itemised menus are created. Finally, the dietician nutritionist reviews the automatic recommendation and triggers the relevant restrictions or fine-tunes the algorithm’s suggestion. The project is funded by the Spanish Ministry of Science, Innovation and Universities as part of the Challenges Programme.

• The robot that keeps you company

Robotics is still very much experimental in Catalonia but we will increasingly get used to seeing it in healthcare, nursing homes, day-care centres and schools.

In Catalonia, the Pepper robot went around hospital wards in 2018. We included it in this social section (and not in the healthcare section) as the machine’s main functions were to inform patients or keep them company. It moves silently and has a humanoid appearance and can also interact with people in 21 languages. It is still in the development stage but it has already been designed to explain to older people how they should take care of themselves or how they should treat their illness. It would also be helpful in supporting immunocompromised children who have to be kept in isolation after surgery.

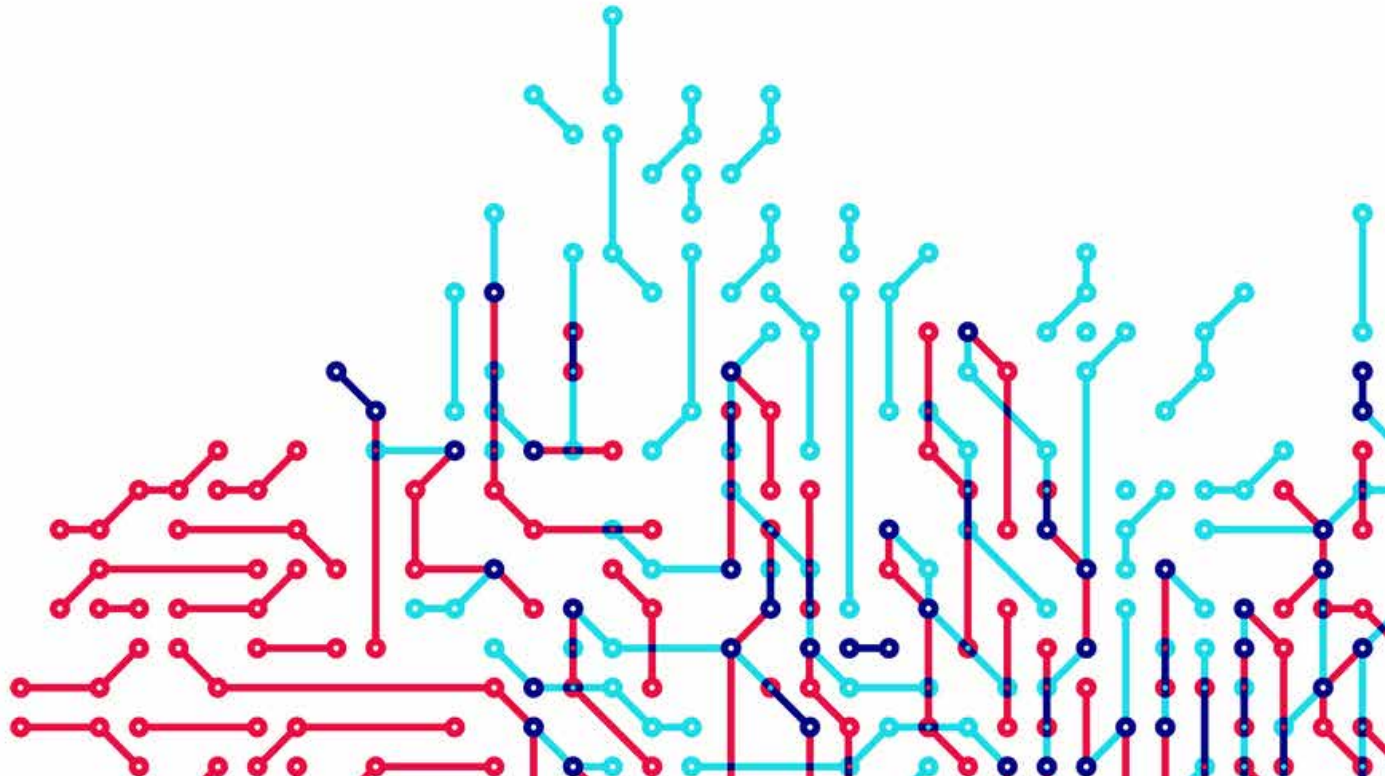


The robot draws on machine learning techniques and can recognise people very accurately by memorising their facial features. It can also identify patients' moods based on facial expression and tone of voice. A number of medical facilities have teamed up to carry out the project including Sant Joan de Déu and Clínic hospitals, the Robotics Department at La Salle University and the firm YASYT.[64]

• Gavius, to be more social

A project that is kicking off in 2020 at Gavà Town Council is Gavius.[65] A virtual assistant will tell the public what social assistance benefits are available, how they are processed and awarded and how they can be conveniently, quickly and easily received - all by mobile phone.

With €4.3 million in funding from the European Urban Innovation Actions fund, it is a joint project between government (the AOC Consortium, Gavà and Mataró town councils), business (GFI and EY), the public (Xnet) and research facilities (UPC and CIMNE).



64 YASYT (https://yasyt.com/ca).
65 Gavius (https://www.aoc.cat/2019/1000265639/laoc-collabora-amb-gavius-un-projecte-innovador-en-lambit-dels-ajuts-socials/).

2.4.7. Employment

Technological innovation in the staff recruitment industry is nothing new: the first test forms appeared in the 1940s, digital techniques were already being used in the 1990s, and now it is the turn of artificial intelligence. And it is possible to reach previously unimaginable levels. For example, the Finnish recruitment agency DigitalMinds has about twenty large corporations as clients. Since it receives hundreds of CVs from candidates, it seeks to save time in selecting the best employee for its clients and therefore does not conduct personal interviews: instead it asks for the applicant's email address (and social media profiles) and an automated decision-making algorithm scans all the personal information (messages sent and received, interactions, etc.) in addition to deciding whether they should get the job.

In Catalonia there are also AI applications in the employment market. Fortunately, they are not as intrusive as the Finnish case. AI is already used for automating staff selection or predicting the probability that an unemployed person will find a new job.

• Staff selection by facial gestures

The Computer Vision Centre presented a demo of a human resources recommender at the Mobile World Congress 2019. There is a research group which based on behavioural and gestural analysis of people can tell what kind of occupational profiles and which skills candidates have most and least developed.

Let's picture a person going for a job interview. The human resources director does the interview but there is a camera recording it and software equipped with algorithms that are capturing personality traits.[66] "The human resources manager will rely on their personal impression but also on the analysis by the smart system to decide whether this person is the candidate they are looking for or not," explains Meritxell Bassolas, Director of Knowledge and Technology Transfer at the CVC. "The machine can detect whether their behaviour is nervous, confident, less artistic, more reflective, etc. And it does so through gestures as well as through facial emotions and microexpressions." However, she thinks we will never have interviews solely with a camera and assessments by algorithms: "These systems should always just be support for the professional's decision. Human resources experts have a lot of knowledge that they also need to bring to the table when recruiting people. They know what parameters they have to evaluate in line with the post they are looking to fill or the skills needed. The machine does not get this far."

66 CVC at Mobile World Congress 2019 (http://www.cvc.uab.es/outreach/?p=1780).

• The biases of professional platforms

The Web Science and Social Computing Research Group at the UPF teamed up with the Technical University of Berlin and the Eurecat Technology Centre to build an algorithm which identifies and mitigates biases in other algorithms. They called the system FA*IR.[67] "We studied data about job offers, prisoner reoffending and university admission rankings to identify patterns of discrimination in directories that may benefit or disadvantage certain groups by gender, age or race," says Carlos Castillo, Director of the Web Science and Social Computing Research Group at the UPF.

One of his PhD students, Meike Zehlike, investigated how men and women were classified on the professional platforms LinkedIn and Viadeo. "If there are 100 profiles of equally qualified men and women and only men appear in the top results of the search engine, then we have a problem."

FA*IR detects this type of discrimination and corrects it by adding an affirmative action mechanism to reorder the results and avoid discrimination without affecting the validity of the result.

• Predicting sick leave

Based on the attendance or absence of staff in a hospital, an algorithm can predict how many cases of sick leave can be expected every day in each service and job profile. The system provides an estimate of how many people will not turn up for work.

"We never profile workers; we do it by services," explains Ricard Gavaldà, coordinator of the research laboratory at the Technical University of Catalonia (UPC). Nowadays health personnel are hired on a daily or weekend basis and they are assigned to A&E or to wards, but they don't have the experience of having worked regularly at that facility. This leads to distortions, there is no time to train the professionals, they do not know the environment of the hospital, etc. "You can't predict the reason why sick leave will be taken, but you can predict that the service will be short-handed on certain days," argues Gavaldà.

• Automating to manage in record time

One of the roles of the Catalan Public Employment Service (SOC) is training working or unemployed people. Handling the funding for this training is a huge time and resourc-

67 M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, R. Baeza-Yates. "FA*IR: A Fair Top-k Ranking Algorithm" (https://arxiv.org/pdf/1706.06368.pdf).

es problem because there are lots of public and private organisations delivering public services which apply to deliver the training and get money from the government.

"We're talking about large sums of cash which might be up to €50 million, and 300 or 400 training institutions might apply. Each file is a project and each project involves three or four courses," say the SOC's Technical Secretariat staff. "Before awarding the funding, internal staff had to enter all the variables by hand in order to evaluate them based on preset criteria. And so on, file by file." The enormous volume meant it was months before the organisations which would provide the training were known.

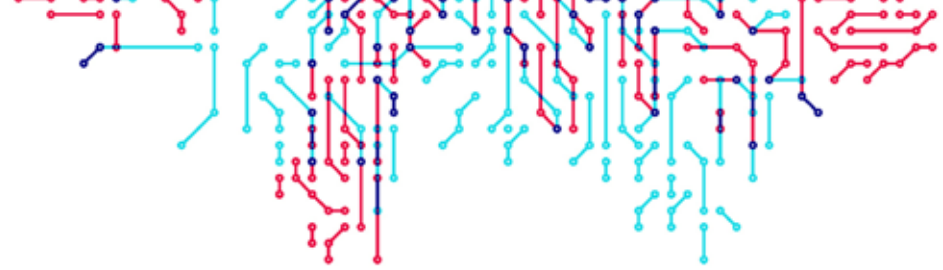
In 2014, it was decided to automate this process and now it is done in just a few hours. In this case there is no machine learning, but there are logical command algorithms which are applied to a machine. There is a logical sequence of varying complexity because it affects many parameters, which takes care of a burdensome and complicated task.

• Automating to plan better

The Eurecat Technology Centre has designed an algorithm which analyses the offers on the Infojobs and Feina Activa[68] job portals (the ones most frequently viewed by employers). The aim of this pilot project is to show that artificial intelligence can be extremely handy for classifying training options by geography and by industry. "We can find out the shortcomings, the overlaps, and also match the skills asked for by employers with the training by the relevant departments, the Catalan Public Employment Service and the Ministry of Education," say the SOC Technical Secretariat staff.

One of the potentialities is to support careers advisers. Depending on the person's characteristics and employment experience, the algorithm could predict what the likelihood of finding a job is. "For example, if you are a woman with a university education who has held a number of management roles, you are classified in a certain cluster. Using the previously introduced big data of the whole population of Catalonia, the algorithm will be able to predict the percentage probability of finding a job." Next the SOC careers adviser along with other career guidance officers will suggest the options that best fit your profile to improve your employability. "This is another tool the adviser has to hand; automated decision-making is never used to award or deny training or any other service as it might be discriminatory." This scheme has just been started up in 2020.

68 Catalan Government's Feina Activa portal (https://feinaactiva.gencat.cat/web/guest/home).



2.4.8. Cybersecurity

Computer attacks on businesses whether large and small and on governments and organisations which handle large volumes of private or financial information call for increasingly sophisticated cybersecurity. “Companies engaged in cybersecurity use algorithms to describe the normal operation of all the computers and interactions which take place in the organisation’s environment based on information found in a range of data sources,” says Manel Medina,[69] professor at the Technical University of Catalonia and founder and director of esCERT-UPC, the Spanish computer security incident response team. “They correlate them and can thus identify deviations from the norm which are indicators of a potential anomaly.”

But cybercrime is growing and as most of the keynote speakers at the Barcelona Cybersecurity Congress held in October 2019 acknowledged, people who want to do harm have an easier time of it than their counterparts in shielding and safeguarding. “Cyberthreats are a growing trend worldwide affecting all industries,” says the Catalan Government’s report *Cybersecurity in Catalonia*.[70]

The main reason for this is the digital transformation which has taken place in all sectors of society over the last decade. “The same technological progress which has driven business productivity and efficiency is what has made organisations more vulnerable to cyber attacks,” the report adds. It notes that every time a company is attacked it is forced “to discontinue its operations for an average of 17 hours a year. Other negative effects may be complete shutdown of operations; decreased turnover; disclosure of confidential information; legal issues; loss of product quality; damage to physical property and even harm to humans.”

The worst thing is that the complexity of the threats is swiftly and steadily increasing. This forces companies to be eternally watchful. Medina explains that today’s automated systems can detect if you connect to IP addresses in unusual countries for the organisation’s activities or if any of the company’s computers have data traffic which may not be normal in order to identify potential information leaks. “The problem comes with advanced persistent threats (APTs);[71] in other words, spyware which auto-installs on company computers and lies dormant for long periods of time. During this time it generates low intensity data leaks and this means anomaly detection software does not perceive significant deviations from normal behaviour.”

69 Manel Medina (https://inlab.fib.upc.edu/es/persones/manel-medina) (https://www.linkedin.com/in/manelmedina/?originalSubdomain=es).

70 *La ciberseguretat a Catalunya* (https://www.accio.gencat.cat/web/.content/bancconeixement/documents/informes_sectorials/ciberseguretat-informe-tecnologic.pdf)

71 “APT” (https://en.wikipedia.org/wiki/Advanced_persistent_threat).

APTs are stealthy and continuous processes intended to bypass a company or organisation's IT security, usually for business or political reasons. And they are programmed to remain active for a long period of time. "Four years ago there was a theft of personal records from the United States Office of Personnel Management (OPM) which even included people who had retired. Three months earlier, the government had protected it with an intrusion detection program."[72] There was no explanation of how this could have happened and the reason was that the leakage of information was already taking place when the screening program learned what could be considered normal behaviour. "They are machine learning systems[73] which learn the normal parameters of a supposedly normal dataset and if you give them a sample that does not fit this normality, they tell you," explains Medina. "But if the 'normal' sample already contains the traffic generated by the spyware, it is seen as part of normality and will not trigger an alert."

• The new challenge: biometric data theft

Nowadays human identification usually involves fingerprints, iris scans, facial recognition, voice recognition or DNA. It is increasingly common for companies to compel employees to choose one of these systems to ensure that some workers do not clock in on behalf of others.

Biometric data are unique and non-transferable. No two faces are the same and no two fingerprints are the same. In an article in Forbes magazine[74] a few months ago, journalist Jayshree Pandya pointed out that "the rise of biometric technology and its use in human identification and authentication will likely have a profound impact on human society. While the rapidly evolving biometric technologies seem to offer the much-needed identification and authentication solution for nations, their use is also raising some security concerns. […] nations are simply not prepared to secure the rapidly growing biometric data or indicators. […] Considering the impact that it may have on human society, the risks to performance, accuracy, privacy, interoperability, multimodality, and even potential health risks (vision risks associated with retinal scanners and more) need to be effectively managed."

72 "Intrusion Detection System (IDS)" (https://en.wikipedia.org/wiki/Intrusion_detection_system) (https://en.wikipedia.org/wiki/Office_of_Personnel_Management_data_breach) (https://www.csoonline.com/article/3130682/the-opm-breach-report-a-long-time-coming.html).

73 "Machine learning" (https://en.wikipedia.org/wiki/Machine_learning)

74 "Hacking Our Identity: The Emerging Threats From Biometric Technology" (https://www.forbes.com/sites/cognitiveworld/2019/03/09/hacking-our-identity-the-emerging-threats-from-biometric-technology/#353ed3505682).

Manel Medina at the esCERT-UPC also warns about these security risks which in principle are sold as a technological breakthrough. “Companies should be aware that if they gather biometric data, they need to be closely safeguarded!”

• Algorithms to protect customers

The Internet has changed the way in which goods and services are bought and this has led to a rapid makeover of the internal organisation of companies. If they do not invest in security, the cost of a major computer attack may result in serious financial and reputational losses.

The risk of being the victim of increasingly sophisticated cyberattacks is very high. Unfortunately, small businesses and shops always lag behind in digital transformation and only think about cybersecurity after an attack. “Cybercriminals targeting this sector have developed advanced and often automated TTPs (tactics, techniques and procedures) to compromise and monetize stolen data,” explains the report *Cyberthreat Intelligence for Retail & E-Commerce*[75] published by BlueLiv (one of the cybersecurity start-ups founded in Barcelona and which already has offices in San Francisco and London). “From convincing phishing[76] campaigns tricking users into sharing personal and financial information, to account hijacking to commit fraud, to crimeware and targeted malware attacking PoS, digital payment systems and customer databases, digital risk has never been so high.” Automated decision-making algorithms provide real-time fraud detection notifications for stolen credit cards and they also prevent it by intercepting cards before they are resold on the black market.

• Machine learning to detect attacks

Artificial intelligence is being used for both defence and attack. “It’s a race to see who can defend or attack best,” says Josep Domingo Ferrer,[77] Director of the Centre for Research in Cybersecurity in Catalonia (CYBERCAT) and a professor at Rovira i Virgili University. “Deep learning methods trained with past data are used to detect potential threats. For example, the characteristics of the perpetrator of previous attacks are analysed and a warning is given when the same conditions are met. Similarly, previous cases are analysed to find the virus that led to an attack. The algorithms work very well when large historical databases are available.”

75 *Cyberthreat Intelligence for Retail & E-Commerce* (https://www.blueliv.com/thanks-ecommerce-retail-whitepaper/).
76 “Phishing” (https://en.wikipedia.org/wiki/Phishing).
77 Josep Domingo Ferrer (http://www.urv.cat/es/universidad/conocer/personas/profesorado-destacado/2/josep-domingo-ferrer).

2.4.9. Media and communication

Automated decision-making algorithms have been applied for more than ten years in the media and communication sector, especially with computer vision techniques to interpret sign language, to give an example.[78] This means hearing-impaired people can hold a conversation and communicate with others who do not understand sign language as they translate from sign to word in real time.

More recently, AI has been implemented in Wikipedia to detect vandalism or incorrect entries and also to generate news in media or to assess how series and movies are shown on video-on-demand platforms such as Netflix.

Some ethical dilemmas, red lines that should not be crossed, and the peril of living in filter bubbles due to business interests are raised. There are also questions that you would not even want to ask: will machines replace journalists? Technologically it can already be done. Or this one: could we end up losing local audiovisual production because an algorithm only shows me the American options? Technologically this is already happening.

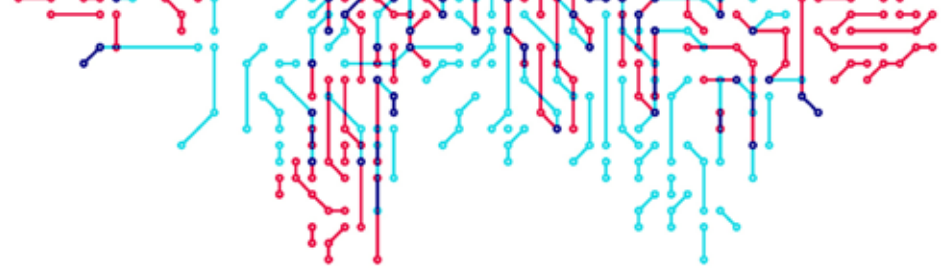
• The journalist algorithm

"One of the most interesting and promising areas of artificial intelligence is machine learning," says David Llorente, founder of a company which generates news using AI. "The machine is really accurate in creating content. It will not say that another player has scored the goal, and nor will it make mistakes in the election results of this or that party or in the weather forecast because it draws on the data."

Llorente's company works for 25 Spanish media and news agencies. Another issue, says Llorente, is social perception. "It is only natural for journalists to feel professionally threatened. But there is still some way to go. A machine will never be allowed to produce a story which accuses people of fraud."

This company's algorithm only writes news stories about topics which have objective data, i.e. election and sports results, financial figures, lotteries, weather forecasts, traffic, etc. "The media outlets which hire us see how we double or triple the volume of news they produce every day and with a high degree of accuracy. This helps them sell more subscriptions or get more advertising."

78 "Creat un sistema visual per interpretar llengües de signes" (https://www.uab.cat/web/noticies/detall-d-una-noticia/creat-un-sistema-visual-per-interpretar-llengues-de-signes-10902264341 00.html?noticiaid=1275458325318).



Llorente explains that the smart system produces the news in three stages: in the first, the machine decides the most relevant data for building the story. “In election results, the most important thing is to say who has won,” he says. “Whereas in a football match, describing a goal when eight have been scored is perhaps not so important.” In the second stage, how to word the information is decided: based on a volume of similar news items, the machine learns the structure of the story and the form. There is also a final review by journalists working in the company to confirm the narrative. Once the system has been trained, it takes the final data, selects the sentences and writes the end story. The machine is always trained using the newspaper library of the media outlet which it is working for to get the writing tone and style. You not only gain in speed and volume of information but also in accuracy and correct spelling.

David Llorente thinks that nowadays journalists do lots of routine jobs which are not really creative and could be entrusted to smart systems. “What we are trying to do is to give journalists more time to produce more thoughtful information, context, in-depth reports, etc.” However, he believes the machine should be prevented from expressing an opinion even though it can already do so technologically. “The opinion issue is trickier. Opinion is generated by humans. You have to be very careful. These are red lines for us. Because right away we could be accused of creating fake news with misleading headlines on big issues like climate change or political topics which might alter the way people see things. The ethical issue is something we are very aware of. We have received offers to do this kind of work and we have turned them down outright.”

• The damned filter bubble

The Catalan Audiovisual Council (CAC) commissioned researcher Carlos Castillo[79] to study how the algorithm presents videos on platforms such as Netflix and OrangeTV. The EU requires 30% of European audiovisual production on each platform. “But it is one thing to have this audiovisual production in the catalogue and another for the algorithm to end up showing it to the user,” says Castillo. “The new review of the Audiovisual Media Services Directive (AVMSD) specifies that they have to guarantee this amount of European production to users in their catalogues. If the algorithm has shown it to you the first few days and you haven’t selected it, it may not show it to you again because it assumes you’re not interested.”

Castillo argues that the first thing we have to bear in mind is that users have the feeling they are choosing the series or movie they want to watch on a VoD platform. “But it’s

79 Carlos Castillo, Distinguished Research Professor in the Department of Information and Communication Technology at Pompeu Fabra University. Report for the CAC: “La oferta y la demanda del contenido audiovisual en la Era de los Datos Masivos”. 2018 (https://chato.cl/papers/castillo_2018_oferta_disponiblidad_contenido_audiovisual-ES.pdf).

not like that at all. Maybe you have forty titles in front of you but this is a very small part of the total catalogue. And the user doesn't decide the criteria by which those first forty titles have been selected."

The report commissioned by the CAC discusses the threats to autonomy and diversity, the information bubble, living in a filter bubble. "If you only watch action movies, it doesn't show you more than that. So there is no diversity here. There is also a threat to autonomy as you should be able to explore the whole catalogue. Now there's only one little search box which makes it impossible for you to access it in its entirety. There could be other interface designs which for example allow you to search by voice. Similarly, they could give you the option of choosing whether the catalogue is sorted or unsorted."

• AI for Wikipedia in Catalan

One of the biggest concerns of open content Wikimedia projects is reviewing potentially harmful contributions ("edits") along with contributions that unintentionally contain errors. Since 2018, Viquipèdia (Wikipedia in Catalan) has used machine learning to review articles through ORES,[80] a tool that "flags up actions which may be vandalism, for example. This means volunteers who review recent changes to Viquipèdia can do their job more easily," they explain on their website. "To enable ORES, the community of editors has to *teach the system* by tagging edits already made, indicating whether they are spelling, grammar, content updates or where applicable vandalism."

80 "ORES" (https://ca.m.wikipedia.org/wiki/Viquiprojecte:Viquirepte/ORES).

2.4.10. Computer vision

Computer vision (CV) is a branch of artificial intelligence which gets algorithms to understand an image just like people do. They can now make out objects and people and also describe them or say what these people do.

CV is not new ground. "It's been more than forty years since research began. In the 1970s, the scientist and father of artificial intelligence Marvin Minsky[81] experimented with it," points out Petia Radeva, director of the Machine Learning and Computer Vision research group at the University of Barcelona (UB). "About thirty years ago, CV was brought into automotive factories to assemble cars and do other jobs in controlled environments. The new thing is machine learning because it can solve more complicated problems such as driving autonomous cars. Research now is twenty times faster than a decade ago."

We decided to include a separate section on this technology because although the examples it contains could have come from the healthcare, social or business fields, it is useful to underscore deep learning's potential in CV.

Where is computer vision used?

It's already all over the place: in car parks when a machine reads the car's number plate and opens the exit barrier; in facial recognition systems using the iris and fingerprints employed by airport security and the police; and also in our mobiles as a stand-in for a password.

CV is in the intercoms of buildings, companies and some houses. Some primary and secondary schools have begun using it to monitor students in their facilities. Banks offer it to their customers to access their current accounts at an ATM. In hospitals, CV is used in the analysis of all kinds of body images such as CT scans, X-rays, densitometry, MRIs, ultrasound scans and mammography to detect and predict diseases.

It is also used to count people, such as the number of demonstrators in a public space, and in sports, for example the number of times each player touches the ball in a match, attacks, fouls, etc. These algorithms are not strictly speaking automated decision-making but rather they calculate, estimate and quantify. "Humans are better at qualifying, machines are better at quantifying," says Petia Radeva.

On mobiles it is useful for adding tourist information when taking a photo of a sight regardless of the angle or type of light at the time. It is also used to recognise scanned

81 "Marvin Minsky" (https://en.wikipedia.org/wiki/Marvin_Minsky).

documents from past centuries and to say what an object found in an archaeological site actually is. Plus CV is in the robots at Amazon's automated plant in Catalonia and in the ones that move along hospital corridors (still in the testing stage) to bring food to patients or keep them company. And at the 2019 Mobile World Congress[82] in Barcelona people entered quite literally face first: instead of showing their accreditation, the barriers could also be opened with facial recognition.

One last example: CV-based iris recognition enabled photographer Steve McCurry to find Sharbat Gula, the Afghan woman photographed in 1985 when she was only 12 years old and fleeing her country at war. Her face became world famous because it was published on the cover of *National Geographic.*[83]

Below are some examples that can be found in Catalan factories and industries and which are the outcome of the teamwork of a number of research centres.

• Detecting defective pizzas

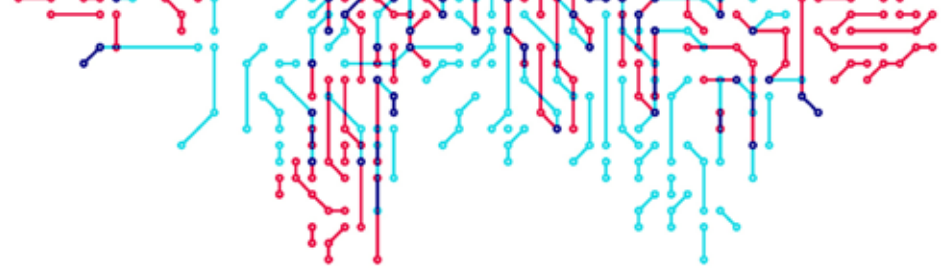
The Computer Vision Centre (CVC) at the Autonomous University of Barcelona (UAB) has been helping Catalan businesses to implement digital projects and industrial processes for more than fifteen years. "When we talk about artificial intelligence, we think of autonomous mobility and large-scale algorithm decision-making," says Mertixell Bassolas, the Director of Knowledge and Technology Transfer at the CVC. "Yet there are many other exciting, smaller scale projects which are driving major breakthroughs."

"In the food industry and with deep learning, we can identify whether the appearance of the pizzas is right or not, whether the ingredients are laid out properly on the pizza crust and whether there are intrusive materials in the packaging," explains Bassolas. "To do this, the algorithm ideally needs to be trained with lots of images, millions; the more images, the more accurate the result. And also of pizzas in good and bad condition if possible, so that it learns to tell them apart. And it needs notes to decide what each ingredient is."

Bassolas says the challenge is to get what we want but with minimum initial effort on our part. Videogames are used to generate synthetic images of virtual pizzas or autonomous cars, as many as the algorithm needs to be trained. "But this is also a lot of work," she adds. "And in the end we get a smart system with very little information which can re-

82 "Comença el Mobile World Congress amb la mirada posada en el 5G". Betevé. February 2019. (https://beteve.cat/economia/comenca-mobile-world-congress-2019-5g/).

83 Sharbat Gula (https://www.nationalgeographic.com/news/2017/12/afghan-girl-home-afghanistan/).



solve what we want. And one which also does not forget what it has learned so it can be used in another case without starting from scratch."

• Monitoring pig fattening

Sectors such as the food industry are slowly going digital. Processes that until now were done very much by hand are being automated in order to draw new conclusions which help to get more out of resources or improve products. In the case of livestock farms, the entire process is already monitored: from fattening to the animal's arrival at the slaughterhouse. "We monitor the pig throughout the entire value chain, which allows us to compare what type of feed is best both to shorten times and also to enhance the quality of the end product," says Bassolas.

The Computer Vision Centre designed a tool to identify the pigs at the slaughterhouse. "We're using neural networks that today can already detect any movement in a scene." The same smart system is also implemented in the fattening farms to track the pig's evolution, its weight, volume and condition and to assess how this impacts the quality of the product.

This system had previously been used to sort waste at a rubbish plant and to read numbers on water meters.

• The complexity of autonomous cars

We have heard a lot about the wonders which autonomous cars will perform, yet there is also a lot of uncertainty about how they respond to unforeseen events. Who will be accountable if there is an accident?

Training algorithms for use in mobility is one of the most complicated challenges artificial intelligence researchers have faced thus far. That's because the machine needs to know everything there is around it as it drives. It needs to have previously identified it and recognise it. This ranges from stationary objects (buildings or street furniture) to anything that moves: people and animals and also plants, a sheet of paper, a kite or any object which flies or is thrown by someone, such as a ball. And finally, the car also has to know how to react in every situation. "It is not feasible to train the algorithm in a real environment," notes Bassolas. "So we generate videogames and smart systems are trained using virtual situations."

But after the machine is trained in a virtual environment, will it adapt to the streets of a city or a real road? Until very recently there was no answer to this question says Basso-

las. Now people talk about *domain adaptation*,[84] in other words the algorithm's ability to switch environments and adjust to them. "In an international project we have created a virtual super-environment in an open-source model in partnership with Intel. Companies like Toyota have expressed an interest. Now there are many other brands that are leveraging what we have done to adapt different parts of the process."

• Facial recognition

One of the big problems in facial recognition up to now is that algorithms have been trained with little data from a certain types of people. For example, if they have been used in Europe, perhaps they have been given fewer face or body images of black, Asian or Hispanic people. If this smart system then has to be adapted to a particular situation, it may lead to problems such as the hand dryer in a public toilet which only dries white-skinned hands as it does not recognise the hands of people of colour.

In recent years, the biases of facial recognition algorithms have been widely identified and reported because they are discriminatory, especially against non-white and female faces. In 2015, Google Photos tagged some black people as gorillas.[85] The solution provided was to remove certain tags, such as the term *gorilla*, instead of fixing the algorithm which remains biased. Also, in 2018 Amazon's Rekognition program wrongly identified 28 of the 435 members of the US Congress as criminals.[86]

In Catalonia, facial recognition is used in the judicial system to recognise a potential criminal and also to learn whether the accused is lying in child custody cases. "If the person is lying, the system detects it from their facial microexpressions," explains Bassolas. The question is: can the automated decision-making algorithm get it wrong? Bassolas shakes her head: "This technique has been used and refined for about ten years."

• Satellite photos

Satellite images can help find a solution to a lot of socio-economic problems a country or people need to address. They include cost and resource savings in food production and distribution, energy issues, understanding what has been grown in a country in a year, pollution concerns, natural disasters, floods, etc. "The problem is that the technology used hitherto is very expensive," says Marco Bressan, a data scientist and expert

84 "Domain Adaptation" (https://en.wikipedia.org/wiki/Domain_adaptation).
85 "Google apologizes after Photos app tags two black people as gorillas" (https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas).
86 "MIT researchers: Amazon's Rekognition shows gender and ethnic bias (updated)" (https://venturebeat.com/2019/01/24/amazon-rekognition-bias-mit/).

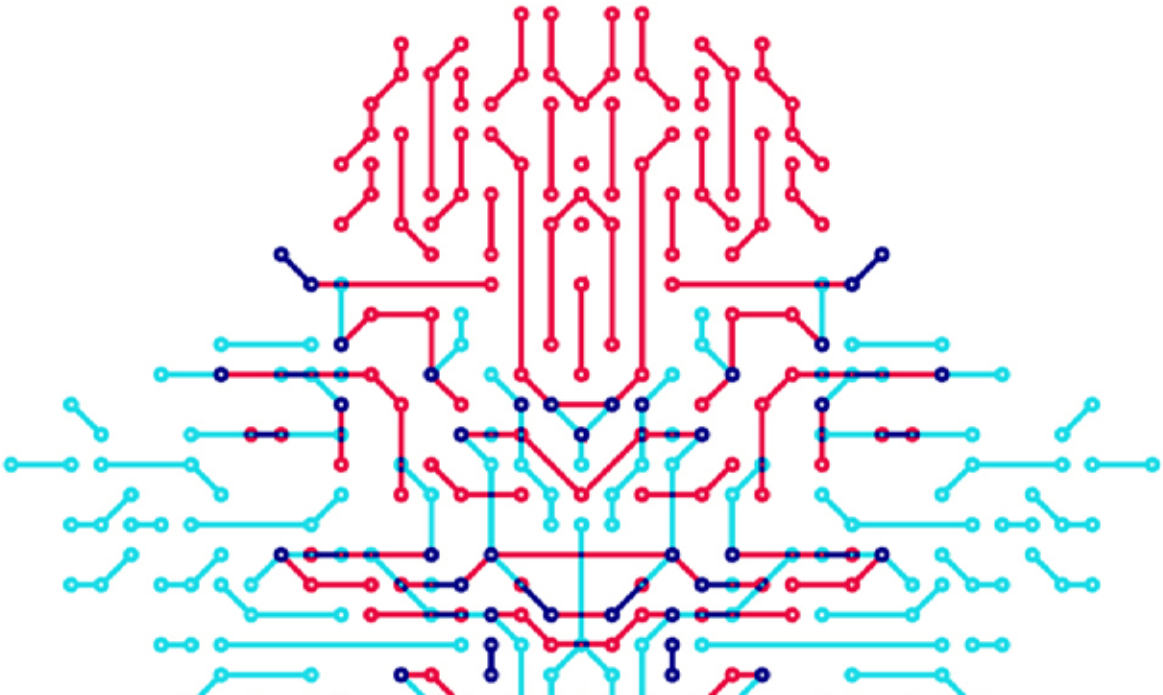
in satellite imaging. “If I want to get a high-resolution satellite photo of the whole of Catalonia updated every month, it might cost me about $400 million per photo.”

“For them to be really useful there has to be a capture frequency in real time. For example, if there is a typhoon in Bangladesh, I want to understand the impact on the rice crops, to have almost daily information on the impact, how many hectares have been flooded, what kind of measures I should take to help people, etc.,” continues Bressan. “To get all these data I have to send not one satellite but hundreds. However, at $400 million a satellite, that’s not feasible.” After some research, the company he leads has managed to maintain image resolution for less than $1 million.

The goal is to get a picture of the Earth every week, and later on one a day. Today, these photographs are useful for the agricultural, forestry, insurance and energy industries. And they are used in particular to manage infrastructures. “If you have a gas pipeline which runs across Siberia or Patagonia, checking the status of the installation for regulatory or safety reasons or because a river has overflowed, trees have fallen down, etc. is very complex and hugely expensive,” says Bressan. “However, a subscription to satellite images of the area of interest can greatly cut the cost.”

“The algorithms analyse everything the images show in real time. If there are cars affected by a flood, how many? In real time you can tell how many hectares are planted in a country. For example, a food producer who depends on this raw material for production can learn about inventory. This was not possible before. Previously, if there was a pest that killed the cereal, you end up paying out a fortune. And now you can predict, etc.”

Governments already use this type of imagery to track refugee camp movements and estimates and in defence for planning drone attacks.



2.5. The ethics of artificial intelligence

The contemporary interest in artificial intelligence is unprecedented. Very soon it will be indispensable for society because it affords a great deal of efficiency, knowledge and creativity. Machine learning, and in particular deep learning, has made it possible to make progress in ways that could not have been imagined a decade ago.

However, you also have to think about the impact it will have on people. It is time to ask questions about the meaning of right and wrong, about the relationship between power and abuse or between bias and distortion. The European Commission's ethical guide for trustworthy AI,[87] the Barcelona Declaration[88] and the AI Strategy for Catalonia,[89] which was presented in 2019 and includes a proposal to set up an Ethics Observatory, are some examples of the concern for safeguarding fundamental rights amidst this technological progress.

The experts interviewed for this report were asked about ethical issues. They all stress the need to find a way to explain how an algorithm has made a decision or predicted a situation and the chance to challenge the automated system's reasoning. They also raise the issue of accountability in the event of malfunction or discrimination by the algorithm. Likewise, they point out the urgency of educating the public including policymakers, educators, parents and public opinion in general. People need to be educated and/or made more aware so that they know when to assert their rights if an automated system has restricted their individual freedom in part or in whole.

87 *EU artificial intelligence ethics checklist ready for testing as new policy recommendations are published* (https://ec.europa.eu/digital-single-market/en/news/eu-artificial-intelligence-ethics-checklist-ready-testing-new-policy-recommendations-are).

88 *Barcelona Declaration for the proper development and usage of artificial intelligence in Europe*. (2017) (https://www.iiia.csic.es/barcelonadeclaration/).

89 *AI Strategy for Catalonia*. Ministry of Digital Policies and Public Administration of the Government of Catalonia (2019). (https://participa.gencat.cat/uploads/decidim/attachment/file/818/Document-Bases-Estrategia-IA-Catalunya.pdf).

Professor of Ethics and Philosophy of Moral and Political Law at the Autonomous University of Barcelona (UAB)

VICTÒRIA CAMPS:[90] "With artificial intelligence we have lost privacy"

What is ethics? And why is it so important to address ethics when using artificial intelligence?

"Ethics could be defined as a set of principles, norms or values that guide our conduct and cannot be overridden by others," says Victòria Camps. "For example, respect for dignity is a value which cannot be got rid of to benefit someone else."

"Religions do not help us to define ethics because while admittedly we have the Ten Commandments in Christianity which tell us 'Thou shalt not kill' and 'Thou shalt not steal', we might have reservations about 'Thou shalt not covet thy neighbour's wife'. Ethics goes further; it includes a demand for universality such as non-discrimination or respect for others."

"Ethics is mandatory but it does not have the force of law, which does compel you and penalises you if you break it," explains Camps. "Ethics compels your conscience. With artificial intelligence there is very little awareness of the dangers. What we used to consider a precious asset has been lost: privacy. Young people (but also people of all ages) have no inhibitions whatsoever and give everything: photos, data, facial recognition. And then there are the videoed rapes, bullying in schools, etc."

Fairness is not achieved by the algorithm alone

"The most important thing in ethics is fairness," says Camps. "A fair system means favouring those who are worst off. The state has an obligation to protect people and their fairness. Can an algorithm do this? Yes, it could favour people who have been evicted, who are unemployed and have children in their care. It could prove that they should be given social benefits. But each case is individual and this also has to be factored in when programming the algorithm. Because there might be people who grumble and the authorities should have answers for the use of artificial intelligence. Fairness is not achieved by the algorithm alone."

"AI cannot skip round the basic questions in ethics about privacy, about confidentiality. You cannot ask the computer scientist who programs the algorithm to think about

90 Victòria Camps (https://www.uab.cat/web/el-departament/victoria-camps-cervera-1260171817458.html).

whether it will be used properly or poorly. The question should be put to the person in charge, to the one who commissions and decides what the algorithm should do. I am a member of the Bioethics Committee of Catalonia.[91] A few years ago there was a lot of discussion in the Catalan Parliament about the marketing of people's health data using the Visc++ program. In the end it was signed off with qualifications and amendments. The question that needs to be asked is this: is having shared medical records progress? Obviously it is. Yet without losing control of how these data may be used." And what would misuse be? "Well, if it becomes known that politician has an unexplained illness. A person has the right to have their health data kept private, and if they consent to their use, only for scientific and medical purposes."

The algorithm's accountability is a recurring topic, especially when thinking about potential accidents involving autonomous cars (ones with no driver) which are bound to come along. "The ethical problems in this case are textbook. Forget about AI. Think about a train driver: they see someone trying to commit suicide on the track. What do they do? Try to avoid the person thus endangering the lives of many more passengers? From an ethical standpoint there's no single answer. You always have to try to achieve the lesser evil. And with artificial intelligence you have to do the same."

Difference between business interests and ethics

Camps thinks that a line has to be drawn between what may be a business interest and what may be ethical. Is it ethical for a company to track our browsing trail and bombard us with advertising in our mailbox, on our mobile, on the websites we visit? Is it ethical for banks to know everything about us, our spending patterns, leisure outgoings and preferences, and seek to build our loyalty with product offers based on their analysis of how profitable we are? Is it ethical for a hotel room to cost me more money than the next person because they know which neighbourhood I live in and can interpret my income?

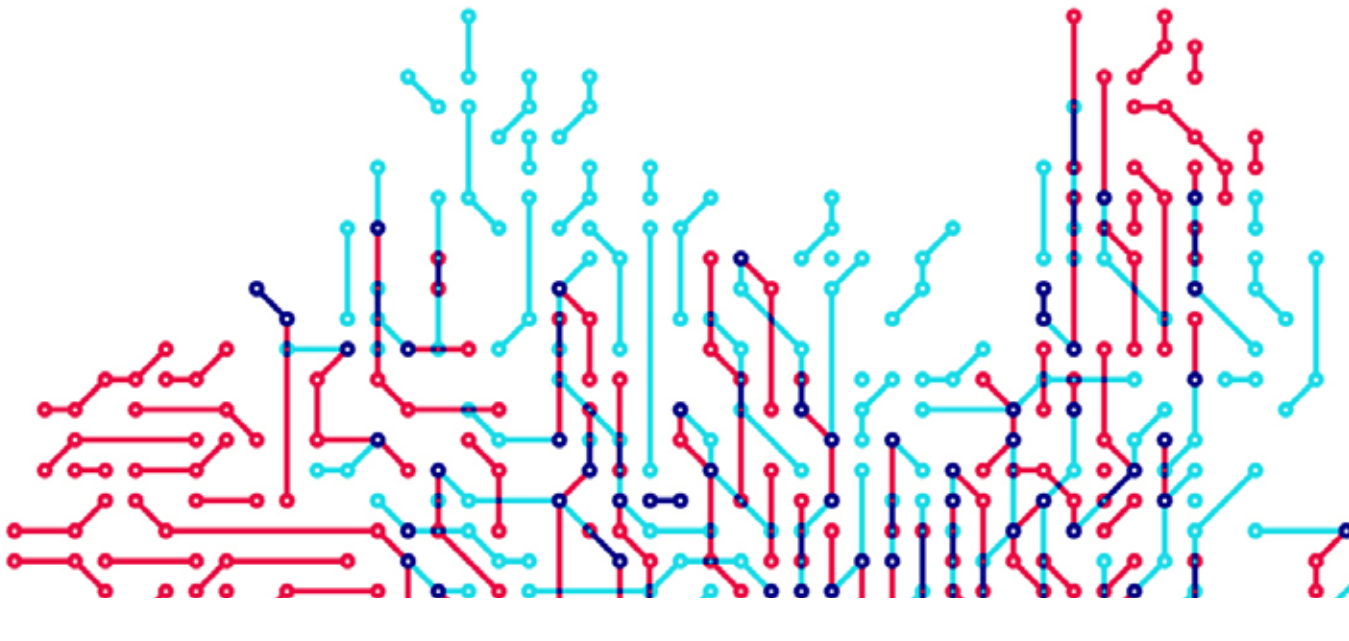
"We live in a consumer society based on supply and demand," says Camps. "And if you buy a vacuum cleaner, all the information or advertising which will be sent to you in the following months may be inconvenient, but that's business. Don't accept it. Another thing is when they want to harm you with your information. For example, the bank may know you are an alcoholic and tells your car insurer so that the insurer can discriminate against you in your premiums."

91 Bioethics Committee of Catalonia (http://canalsalut.gencat.cat/ca/sistema-de-salut/comite-de-bioetica-de-catalunya/).

We must never lose trust

“Life together needs trust in each other. We must never lose it,” says Camps. “You give up some of your freedom in exchange for protection by the political leaders you have chosen. But if you entrust your money to politicians and then cases of corruption come to light, you lose your trust.”

As for the trust we place in social media sites such as Facebook and which constantly let us down because the owner sells personal data to consultants (Cambridge Analytica) which then leverage it to change the way people vote in an election (USA, Brexit, Brazil, etc.), Camps argues that “these platforms are totally superfluous. It’s not the Twitter algorithm that shows you life in a bubble; it’s you because you accept it. What is a pity is that politicians talk to the public via Twitter because that makes politics worse.” And she says that the same thing happened to us before algorithms when we chose our news source media. “But this doesn’t have much to do with ethics. What is ethically important is to preserve individual freedom, so that no one ends up telling you how to live your life.”



Council of Europe rapporteur on Artificial Intelligence and Data Protection. Associate Professor of Law at the University of Turin

ALESSANDRO MANTELERO:[92] "There is no global ethics which can be applied to all countries"

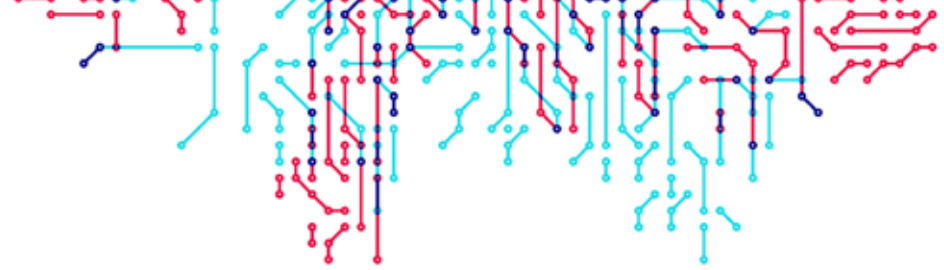
"Ethics and artificial intelligence (AI) is a very hot topic in Europe," says the author of reports on regulation of big data and AI.[93] Alessandro Mantelero argues that when we talk about ethics, we should make it clear what we are referring to. "Ethics can be seen as an additional and complementary layer to the layer of protection of the law. Ethics and law are connected but they are different. If ethics becomes law it has sanctions and it is no longer ethics," he points out. "When we talk about data protection, we are not talking about risks and rewards but about fundamental rights. And we have to bear in mind that ethics in Spain is not the same as in Russia." So then what prevails? Is global ethics possible?

There is no precise answer: "Ethics can never be global; it is up to an environment, to a community. So it is very difficult to get to European ethics. Likewise it won't be the same depending on what artificial intelligence is used for. Ethics will not be the same for an algorithm helping patients as for an elderly care robot. Nor will it be the same for a robot which takes care of a child and has to get particular life values across to them. Plus each family has different values."

Given this complex picture, Mantelero concludes that all the documents and reports currently being drafted will not define a single framework for ethics. "We are at the outset. There is a lot of talk about ethics but little about society. If I say I will invest a lot of money to predict and control crime based on a police application featuring AI, the ethical and social question might be: Why not invest that money to build schools with a good educational system which seek to reduce crime and social difference? This is the way to attack the root of the problem."

92 Alessandro Mantelero (http://staff.polito.it/alessandro.mantelero/).

93 *Regulating big data. The guidelines of the Council of Europe in the context of the European data protection framework* (http://isiarticles.com/bundles/Article/pre/pdf/106203.pdf). *Artificial Intelligence and Data Protection: Challenges and Possible Remedies (https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6).* The European Commission's High-Level Expert Group on Artificial Intelligence is also working on these issues. (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).



Lecturer in the Department of Medicine at the University of Barcelona (UB). Deputy Director of the Bioethics and Law Observatory at the UB

ITZIAR DE LECUONA:[94] "We need to rein in our technological zeal"

The expert in the ethical and legal aspects of biomedicine warns that in biomedical research many research projects which are being evaluated are based on the convergence of technologies, including artificial intelligence and the application of big data analytics.

"Artificial intelligence is already here and we are all contributing our personal datasets to its development," says de Lecuona. "Today we can mine data on a mass scale to improve decision-making by developing algorithms which correlate the data to determine behaviour patterns and predict behaviour. And everyone wants to do this, from businesses to the public health system; to have more efficient health systems or for people to benefit from personalised medicine. Not doing so would be extremely serious from an ethical point of view."

The role of the public

Artificial intelligence is built into the gadgets we carry around with us such as mobile phones. We continuously transmit data from the digital devices we use. Plus there are more complex systems which use our personal data for a variety of purposes.

"The problem is that decisions are made about me using data that is mine, but I will never know who has them or why they are being used," argues de Lecuona. "If ethics is about happiness and freer societies, using AI should make us think about the role individuals play as data subjects. Who is in control and who should be? It's about rethinking the ability to control our data and our privacy in a digital society."

Digital literacy needs to be promoted so that we can make free and informed decisions as citizens. Here de Lecuona points out that there are lots of professionals and companies that team up with scientific research as third parties. They are therefore accountable for programming the algorithms and handling confidential data. In most cases they are professionals who have expertise but have not been trained in ethics.

De Lecuona suggests reviewing which profiles, such as data scientists, should be added to the evaluation of artificial intelligence-based biomedical research by Research Ethics Committees. She also proposes examining what knowledge and training in ethics and

94 Itziar de Lecuona (http://www.bioeticayderecho.ub.edu/es/itziar-de-lecuona).

safeguarding personal data confidentiality should be required for all the stakeholders working in AI to ensure that the principles of respect for people, beneficence, fairness and explainability are met.[95]

Technological laziness

De Lecuona points out that in Catalonia we have some quality health databases including the Shared Electronic Medical Record (HC3)[96] and the Information System for Research in Primary Care (SIDIAP).[97] The challenge now is to make sure the information is interoperable and can be reused. However, de Lecuona emphasises that "we need to rein in our technological zeal and not blindly trust AI in health. The final decision in biomedicine has to stay with the professional, not the machine. Algorithms per se cannot have decision-making power."

From the individual point of view de Lecuona says that we need to shake off our technological laziness and find out about how the automated systems which impact our daily lives actually work. Algorithms should not help to promote inequalities and discrimination or perpetuate them. "We cannot go on thinking that we and our datasets are unimportant and that we don't care who might access them. We are relevant and from time to time we contribute to generate ontologies[98] to design and improve artificial intelligence through our datasets. We need to be careful and foster a culture of respect for privacy. In the digital society, protecting personal data means protecting people."

95 *Principles laid down by the* European Commission's High-Level Expert Group on Artificial Intelligence (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

96 Shared Electronic Medical Record (http://salutweb.gencat.cat/ca/ambits_actuacio/linies_dactuacio/tecnologies_informacio_i_comunicacio/historia_clinica_compartida/).

97 SIDIAP (https://www.sidiap.org/index.php/es).

98 "Ontology" (https://en.wikipedia.org/wiki/Ontology).

Research Professor at the Spanish National Research Council (CSIC). Former director of the Artificial Intelligence Research Institute (IIIA)

RAMÓN LÓPEZ DE MÁNTARAS:[99] "It should be possible to certify that algorithms are bias-free"

Ramon López de Mántaras coordinated the design of the Spanish Strategy for Artificial Intelligence[100] published in March 2019, a major step in setting out the ethical standards for its implementation. He also sponsored the Barcelona Declaration,[101] drafted two years ago by the scientific community and which states very clearly the ethical guidelines that should be followed.

One of the priority points mentioned in the Barcelona Declaration is accountability. In other words, when an algorithm makes decisions, the people affected can get an explanation in understandable terms of why it has made them and be allowed to challenge them with grounded arguments. Yet two years later this is still an unresolved issue which researchers are greatly concerned about.

"As a consumer, as a citizen, you have the right to ask for explanations[102] under the European Data Protection Act. You can ask why you have been refused a benefit or not given a loan. The algorithm's code won't give you one but you should be able to demand that a neutral organisation assesses whether the algorithm is fair. There is no such organisation in Catalonia. Algorithm Watch[103] and the Ethical Tech Society[104] could do this internationally.

"It would be useful if the Catalan government had an agency to certify whether the algorithm is biased or not. This is already done with food and drugs but not with AI. A fairness seal should be set up. Not all algorithms need to be certified, but ones where there is automated decision-making which can significantly harm people should be. This is costly for governments. Yet cars have to do an MOT. If it's done for other things..."

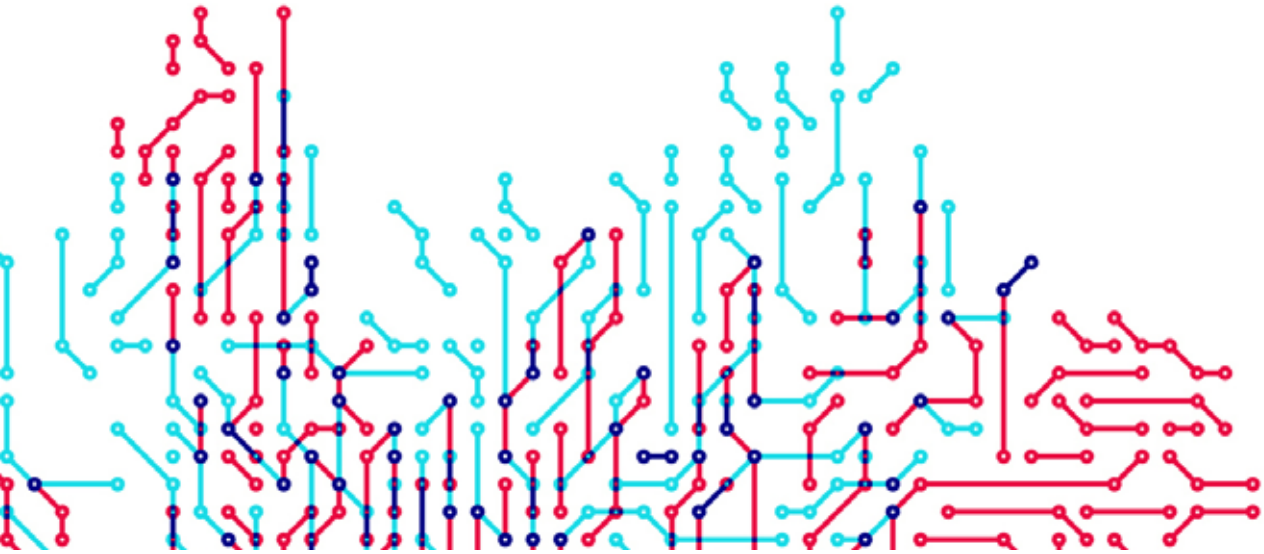
99 Ramon López de Mántaras (http://www.iiia.csic.es/staff/ramon-l%C3%B3pez-de-m%C3%A1ntaras).
100 Spanish Strategy for R&D in Artificial Intelligence (http://www.ciencia.gob.es/stfls/MICINN/Ciencia/Ficheros/Estrategia_Inteligencia_Artificial_IDI.pdf).
101 "Barcelona Declaration for the proper development and usage of artificial intelligence in Europe". 2017 (https://www.iiia.csic.es/barcelonadeclaration/).
102 Ramón López de Mántaras. "La ética en la inteligencia artificial" (https://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/el-multiverso-cuntico-711/tica-en-la-inteligencia-artificial-15492).
103 Algorithm Watch (https://algorithmwatch.org/en/).
104 Lorena Jaume-Palusí (https://twitter.com/lopalasi).

Perpetuating biases

“It often happens that the big data you use to train the algorithm are biased because they are data from the past, and in using them again you perpetuate and enlarge past biases. But you can’t make up the data. You have to get it from somewhere.”

How can you detect bias? “If the decision changes when you remove race information from an algorithm, you have already detected a bias. There is no such thing as being entirely right, but if only the most significant biases could be removed that would be a step forward. Like the soap dispenser which doesn’t squirt out anything when a black hand is placed under it because the sensor only works with white skin. Appalling! And the same with facial recognition that confuses black faces with chimpanzees.”

“The emphasis on the importance of ethical aspects is the main difference Europe can champion compared to what they are doing with AI in the US and China. So the EU requires each country to have a national AI implementation strategy and to clearly set out these minimum standards.”



PhD in Computer Science and Research Professor at the Institute of Robotics and Industrial Computing (CSIC-UPC)

CARME TORRAS:[105] "People who develop the technology should be trained in ethical principles"

"Nowadays a big trend is to work with deep learning, a kind of multi-layered neural network which associates inputs with outputs. It is a type of learning called a *black box*, because we don't know exactly what happens in this association process. If the data are biased, the algorithm learns poorly. But whoever sets it up, whether it's a bank to give you a loan or a public body to evaluate a curriculum or allocate a grant, will not be able to give you an explanation of how that decision was reached. There is a very ambitious European Union project[106] to introduce the concept of *explainability* so as to get explanations about how the machine has processed the data to come to a particular conclusion. The IT community is keen to find a general way of supporting the decisions made by the machine. Plus this support should be given in terms understandable to people who are not computer experts."

Carme Torras has written science fiction novels in which she explores where humans and machines are going. She calls for two things: "1. Regulation to safeguard minimum standards when applying artificial intelligence in society at present. 2. Education, starting with training the people who develop the technology in ethical principles."

Training in technoethics

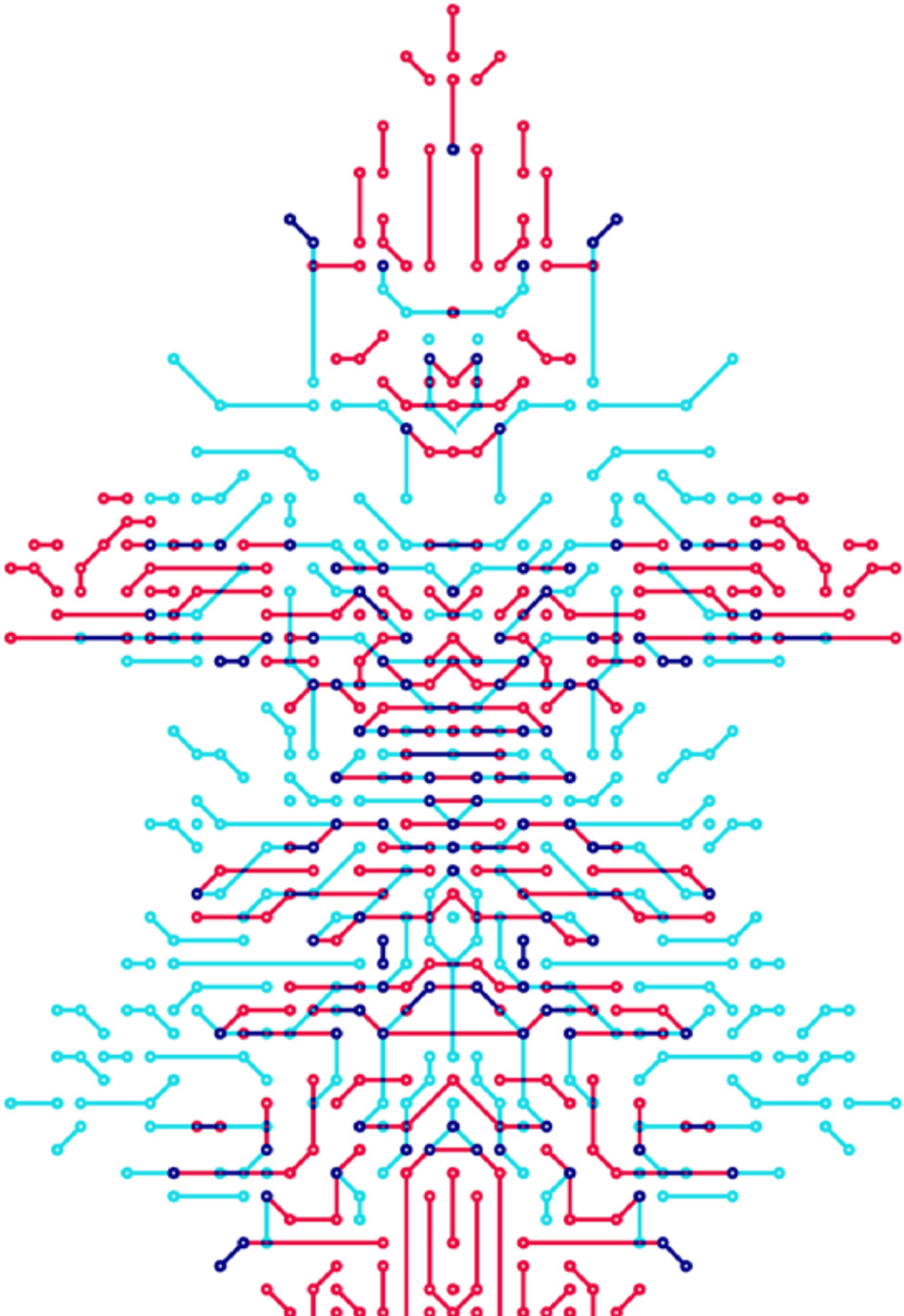
"To take part in this future, you need to know how it is being created," argues Torras. She is an advocate of training trainers so that they can publicise and share the concepts of *technoethics* or *roboethics*, which would be the fusion of technology and/or robotics knowledge with a foundation of ethical concepts.

"I'd like to get to secondary education, to the high schools, to primary school kids. We should start now so it's not too late. This generation will grow up in an entirely technological world which we are building now. It should be an overarching subject in language, mathematics, social studies, philosophy and technology. And it could be called: 'Ethics in the Digital Society'."

She also suggests setting up an information portal which is constantly updated and expanded and where you will find everything from practical tips on safeguarding privacy

105 Carme Torras (https://www.iri.upc.edu/staff/torras).
106 *Ethics Guidelines for Trustworthy AI* (https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1).

on your mobile phone to digital regulations and standards, rights and queries. “I think there is a very large segment in the general public, very diverse people, who would like to learn how to handle things better in this technological era but they cannot find this minimal essential and thorough information service,” says Torras. “We should leverage this interest and give it an outlet.”



CTO at NCENT. Professor of Computer Science at Pompeu Fabra University and Northeastern University

RICARDO BAEZA-YATES:[107] "Today's technological revolution needs ethics"

"Nowadays people generate data in many ways," says Ricardo Baeza-Yates. "And these data are gathered and analysed and have greater value now than in the past. They can be used for legal or business purposes or to tamper with elections. The data already allow us to predict purchasing behaviour or practices. If you told me what you ask a web search engine in a day, I could deduce whether you are male/female, young or old, location, etc. This information that we give willingly along with the trail of our habits can be used for positive or negative ends."[108]

"Today's technological revolution needs ethics," he stresses. "Some regulation, not so strict that it halts technological development, and minimum ethical standards such as explainable AI. In other words, if an algorithm predicts or decides automatically it must be able to tell me how it has done so. For this to happen, people have to know that there are biases in the data that are entered in the algorithms." Baeza-Yates says that technology always moves faster than the social side and that we only worry about ethics when problems arise such as at present.

The Jiminy Cricket of prejudices

If we use the algorithms knowing they have biases, why do we let them make decisions? "Why do we let humans make decisions if they also get things wrong?" replies Baeza-Yates. "We all have prejudices. Automated systems are a great help in situations where biases do not have much influence such as in air traffic control: having people working long hours, under stress, is more dangerous than training machines for this task. They don't get tired, they are programmed and they are more efficient."

"There are a lot of biases; the problem is that we don't notice they are there until they make a mistake. The same goes for our prejudices, which we are not aware of most of the time. Machines can help us realise that and create a fairer world." Okay, but how? "You could create a Jiminy Cricket[109] (or virtual assistant) that would tell you when a prej-

107 Ricardo Baeza-Yates (http://www.baeza.cl/spanish.html).

108 In the Chilean documentary *Por la razón y la ciencia*" *Por la razón* (https://www.youtube.com/watch?time_continue=1253&-v=7PCC7tRyM2I).

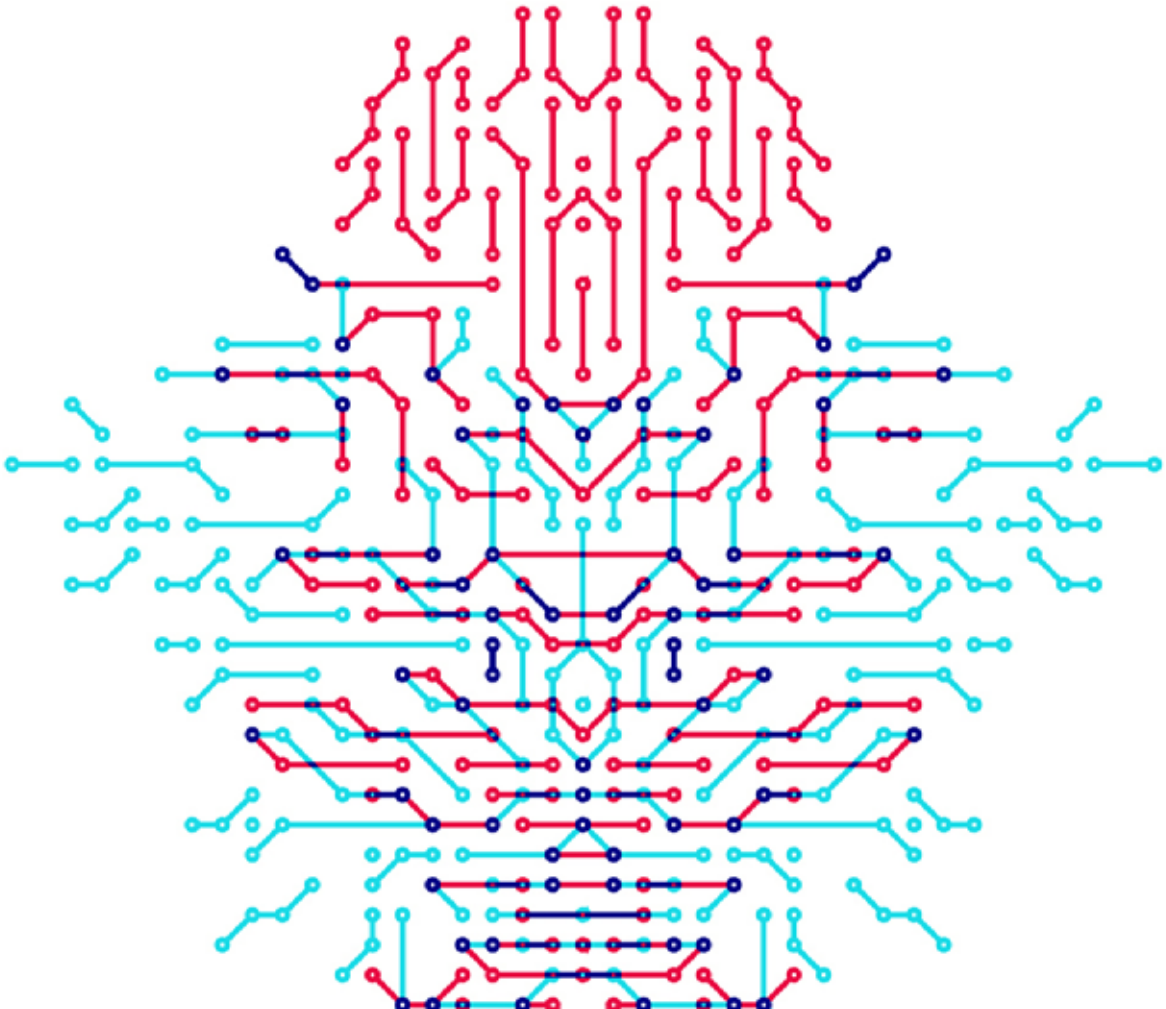
109 in the podcast "La inteligencia artificial tiene que ser nuestro Pepito Grillo". BBVA (https://www.bbva.com/es/podcast-la-inteligencia-artificial-tiene-que-ser-nuestro-pepito-grillo-ricardo-baeza-yates/).

udice is being expressed in speaking, acting, judging, etc. Or to tell us when someone is trying to manipulate us."

And as a follow-up question: how many of us would accept that a machine (a mobile phone or other device) hears what we say (in private and in public) and detects our prejudices?

Improving society

"In the coming years, we will see how identifying biases in algorithms helps to improve the world," says Baeza-Yates to conclude on an upbeat note. But he cautions: "In this new data ecosystem, the trend is to have less and less privacy. When you accept a digital app, you have to read the terms of use and check whether the service you get is worth the privacy you lose. And know whether they can legally ask you for the data they are asking you for. A huge cultural shift is needed to ensure everyone respects the privacy of others." And he recommends more education, being aware that you can easily be manipulated, understanding current technology and upholding rights which safeguard individual freedom.



Scientific Director of the High Performance Artificial Intelligence group at the Barcelona Supercomputing Centre. Professor of Artificial Intelligence at the Technical University of Catalonia (UPC)

ULISES CORTÉS:[110] "We have to teach the limits of technology and critical thinking"

"In the past, it was the algorithms in the banker's or doctor's head that decided to give you a loan or give you heart surgery. You didn't know exactly what criteria they used to make this decision. Now the same thing is happening but with machines. It would be nice if a company could ensure that the decision is explainable if you are unhappy with it. But this is impossible today because no one is designing them to be explainable."

"A small number of people are scared of automated devices because of the mistakes made thus far and in particular the use of machine learning. But most people don't care much about what is done with the algorithms. Yet they should, because their lives are constantly changing as a result of what a lot of machines decide. There have to be some very blatant cases, like Cambridge Analytica, before people start taking this on board. It's not bad to use Facebook if you know the consequences. The bad thing is to let a child sign up without warning them of the risks involved."

"I think it's obvious that in a Western society with a Christian past there is a common set of values. The ones Plato and Aristotle spoke about. In the United States it is worth asking: what is the least harm I may cause? It is quite apparent that what is fairest for us is not fairest for people in Malawi. But we get together and decide what the consensus is."

Ethics means culture

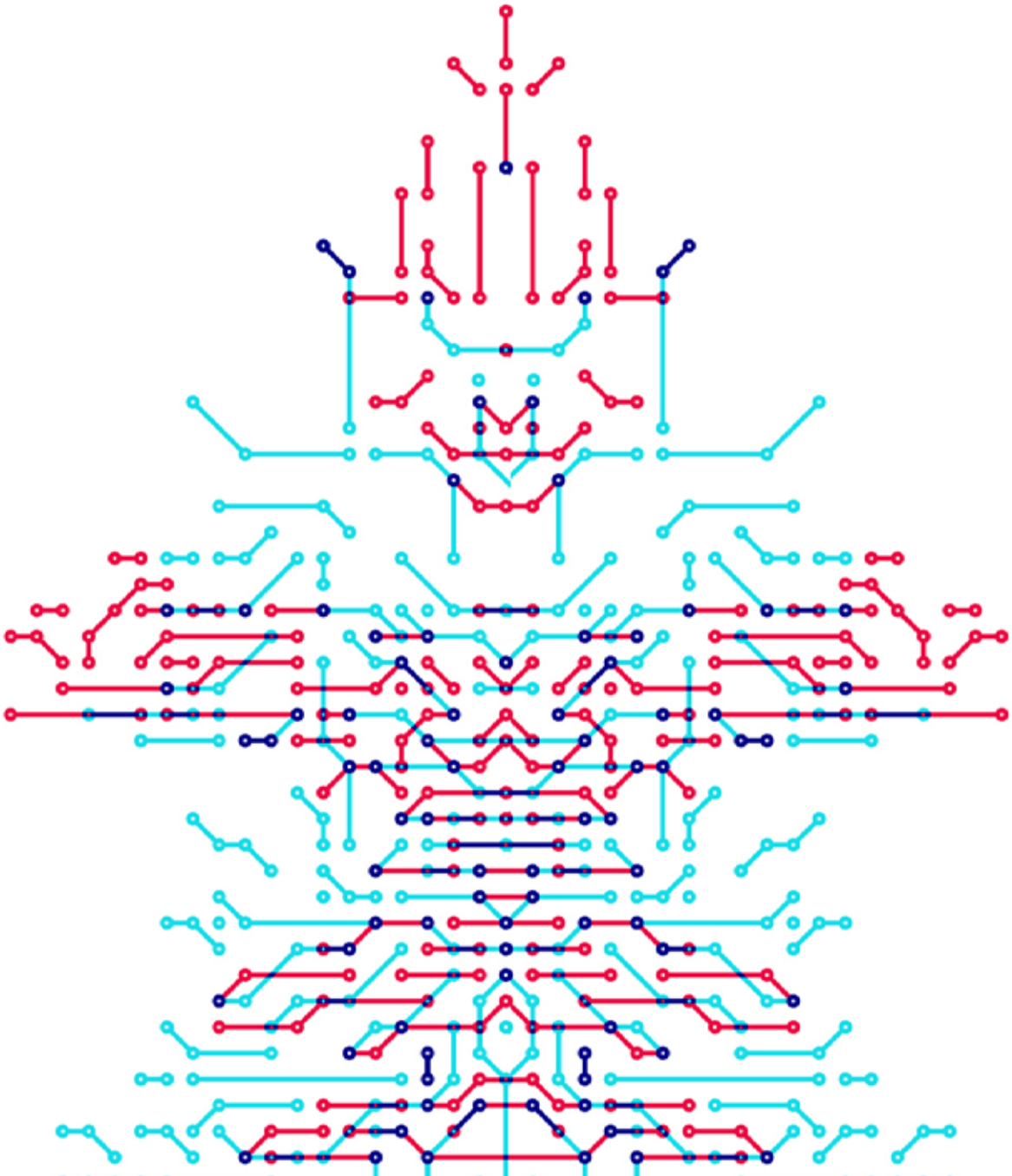
"I would ask everyone three questions: 1. Are you familiar with your country's constitution? 2. Do you know the commandments of your religion? 3. What is the last book on ethics you read? If the answers are negative, how can we ask the engineers to be ethical? You have to put ethics into a culture. And then ask programmers to make algorithms that don't harm the rights of others."

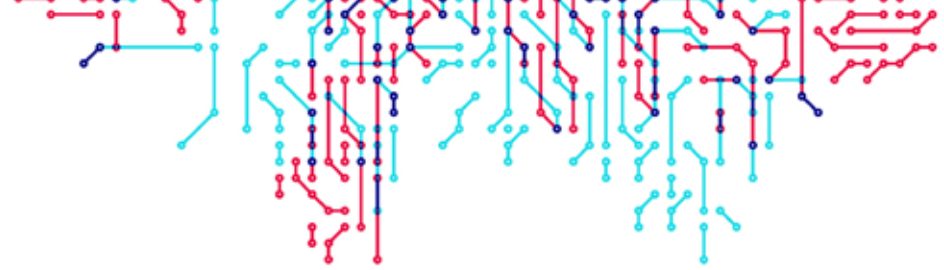
"No technology should be scary. What is scary is the people who use it. Its design is harmless but it can be used for illicit purposes, the bulk of them to make money. Mark Zuckerberg has a great line: 'You can be unethical and still be legal.' We have to teach people to grasp the limits of technology and also critical thinking. People are not interest-

110 Ulises Cortés (https://www.bsc.es/cortes-ulises).

ed in being critical; they are interested in being happy. They live in an incredible dystopia and wonder: 'Why should I worry about my data? I have nothing to hide!'"

"The problem is that we don't even ask ourselves these questions. The only way to turn this situation around would be for politicians to see to the problems arising from technology and to educate educators. Parents have delegated instruction to the state. But when it comes to education about technology, there are no such parents. If you want to change the world, you have to make sure teachers in primary schools, high schools and universities know about technology. And programmers need to learn more ethics. Yet people are increasingly watching TV, spending more time on social media, doing less sport and reading less."





PhD in Computer Engineering and Professor of Artificial Intelligence and Data Science at La Salle-Ramon Llull University.
Member of the Date Science for the Digital Society (DS4DS) research group

ELISABET GOLOBARDES:[111] "Problems are being solved where we don't know how they were solved"

"Everything changes when smartphones out-sell offline mobiles. We democratise the data and now any user can get them and generate them. In Africa they will have gone from having no connection to having 5G very soon so everyone will have a mobile. Everyone wants to know. Power is no longer held by the few; it's held by the people."

"In my computer and robotics undergraduate classes I teach AI-robotics-ethics concepts. It's about education. There is a big hullabaloo with facial recognition too, with the way we unconsciously hand over our unique facial features. We give them to any company which offers us entertainment and we don't know what they will do with them or under what laws. It would be much more worthwhile if we gave our data to a public health service."

"For the first time, technology has gone beyond the mathematical demonstrations that humans do. This is the one dark spot in today's artificial intelligence. Problems are being solved where we don't know how they were solved, especially in certain neural networks in deep learning."

Algorithms are neutral

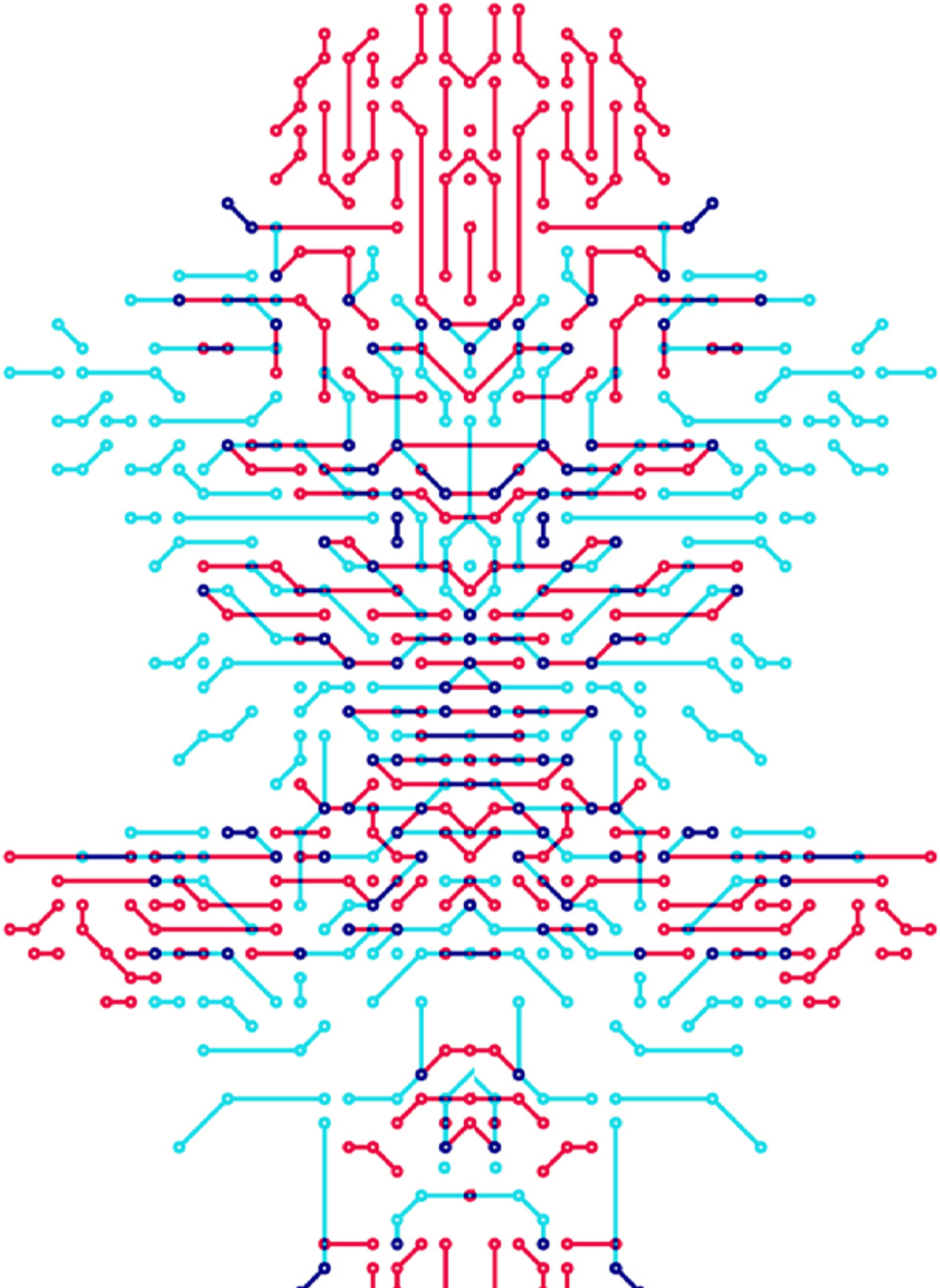
"But I would like to decouple the algorithm from the ethical aspect. The algorithm is a bolt in a car, just a part. It's neutral. The problem lies in the big data you enter to train it and the output target. The programmers aren't to blame either: they do what the company says or whoever commissions the algorithm. It would take public policies to ensure the data for training them are representative of reality, diverse, ethical and egalitarian. The problem is that the person who orders the algorithm to be built for a specific purpose may have a specific financial or business objective."

"An 'Ethical Data' label would be needed, a quality label which would ensure there is no discrimination or bias. Similarly, you would have to make sure that the algorithm's objective has an ethical purpose."

"Nevertheless, I trust artificial intelligence more than some politicians: whose interests are they working for and making decisions about things which impact me as a citizen?

111 Elisabet Golobardes (https://www.salleurl.edu/en/la-salle/directorio/elisabet-golobardes-rib).

They don’t tell me either. I would trust a machine which tries to be fair based on specified criteria more than the unreasonable decisions of many world leaders. Too much attention is being paid to AI, but I don’t know whether it’s to distract us.”



Professor of Computer Languages and Systems at the University of Barcelona (UB). Member of the Department of Mathematics and Computer Science at the UB

JORDI VITRIÀ:[112] "If we don't get our act together, we're in trouble"

"I don't like the word *ethics* because people don't understand it. I prefer to talk about transparency, privacy, diversity and accountability. If a hotel tells me it will share my data with others, I can either agree to it or not. But I know what the game is. If an airline tells me they will share them with the hotel to send me deals, I can accept it or not. This is the *transparency* I want. If a hospital or government agency assures me that they will anonymise my data and use it responsibly, I can accept it or not. But I want this *privacy*. The *diversity* principle must also be in there.

Which company or government can certify that the data do not contain biases which might lead to discrimination? None of them today. Perhaps citizens' associations are needed to champion it. But first you need the interest in and an understanding of how AI works. One of the potential future businesses will be algorithm auditor/certifier to ensure biases are kept to a minimum. Finally, I think *accountability* should be for businesses as well. If I get a Wi-Fi offer shortly after I tell my brother on WhatsApp that I had a problem with my current Wi-Fi, can WhatsApp prove to me that it did not use my data to trigger the offer? Because no one has asked me for my consent to do so. We need to talk about these terms. Ethics by definition is voluntary, but compliance with these four points should be mandatory."

The indirect hazards

"There are other perils for society," says Vitrià. "And in order to address them you can only raise awareness and educate people. One of these dangers is the 'dopamine economy'." He argues that one of today's most valuable assets is engagement, the time that each person spends on a social media site, a game, a medium, browsing a website for buying and selling things, etc. "Online business models try to maximise engagement by offering you what will get you hooked: a series, your followers' tweets, your friends' pictures, the products which will lure you, etc."

Then again, we are very much aware of the facial recognition that countries like China use and of experiments like the one in London capturing the images of all pedestrians

112 Jordi Vitrià (http://www.ub.edu/dept_matinfo/professors/vitria-marca-jordi/).

without telling them.[113] "Yet this year in Barcelona, at the Mobile World Congress, many delegates entered face first. Literally! They were encouraged to enter with facial recognition. If they suggest it to me and even if it's optional, I need to know what the organisers will use all these faces for. Because they haven't explained it and I expect transparency. I think that this is what educating people in technology means."

"In technology they usually tell you only part of the truth. Any technological revolution ends up creating more jobs. But what they don't say is that there is a period of time which may be shorter or longer during which many people are out of work because they will never be able to get into the new professions. You can't turn a baker or a taxi driver into a data analyst. How long would you give taxi drivers? And GPs? There will always be medical researchers, but what about surgeons? They'll vanish without a doubt. Nurses shouldn't worry because we're talking about empathy. If we want to be smarter than in the past, we should design a strategy which enables us to get through the transition period with regulatory mechanisms."

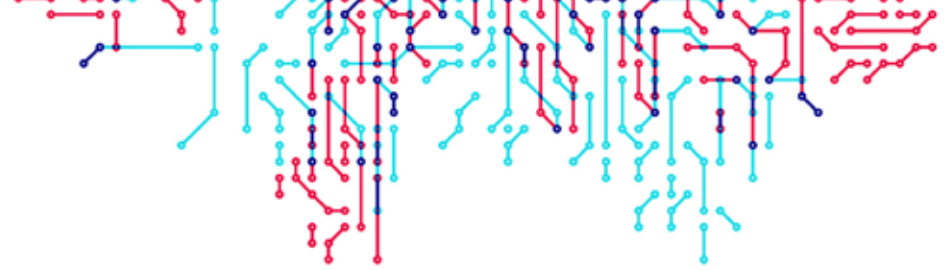
Data literacy

"The problem is that everyone happily uploads their faces on Facebook. Mark Zuckerberg's platform shows an impressive appetite for manipulation. It showed us that in the US election. We still need a lot of education."

"Hospital ethics committees need to understand what we're talking about. The banks? Same thing. And politicians should also be trained in artificial intelligence because they will have to take lots of decisions about it. Ideally a data literacy course should be added to all university courses (literature, science, politics, law or art). The University of Berkeley[114] started doing it in 2017. It makes sense! If you don't have the foundation, you can make a lot of mistakes! Similar projects are also being run in Finland, Australia and New Zealand. The Catalan Government should think about doing it in primary schools, high schools and universities. If we don't get our act together, we're in trouble. We cannot just heedlessly give away our data. We can't believe everything they sell us."

113 Karma Peiró, Ricardo Baeza-Yates. "Algoritmo, yo también existo" (https://www.karmapeiro.com/2019/09/24/algoritmo-yo-tambien-existo/).

114 The message from Berkeley University is: "We live in a world surrounded by increasingly complex data. This course will enable you to answer questions and explore problems you will have to deal with in the immediate future in your professional or private life" (https://data.berkeley.edu/news/fall-milestones-data-science-education).



Lecturer in the Department of Mathematics and Computer Science at the University of Barcelona (UB). Director of the Machine Learning and Computer Vision consolidated research group at the UB.

PETIA RADEVA:[115] "There should be data donors"

Petia Radeva is an international researcher who has been working in healthcare and computer vision for 26 years. She is a staunch champion of technology, machine learning algorithms, computer vision and artificial intelligence.

"The world is going digital. In the last five years we've created more information than in thousands of years of civilization. Algorithms are increasingly crucial in scientific progress. We need them to measure and compare and to analyse whether or not there is a disease, whether an organ is damaged or healthy. Machine learning means machines can learn from training with a dataset. For example, we can show them a thousand cases of diseased organs and the same number of healthy ones. And so the algorithms can build the rules to decide whether an organ is healthy."

Radeva doesn't see a problem in algorithms behaving like black boxes but rather a challenge. "They are admittedly hard to explain. And this is nothing new. Neural networks boomed in the 1970s and at that time people were already accusing them of being black boxes. In practice, computers were not as powerful then as they are now. Real problems were difficult to solve and even more difficult to explain. Now they are extremely useful due to the computing power we have and the amount of data we use to train them. Who could afford to do without them? We not only need powerful and efficient algorithms but also self-explanatory ones. Doctors cannot use algorithms in clinical practise if they do not understand them. So the scientific community has lately been making great strides in developing new methods with the intention that algorithms can generate self-explanations."

Data donors

"Today we can gather loads of data. We have to ask what they are used for. As always, it might be good or bad. I think the challenge of this technological era is to find a way to get everyone to give personal data to science. I do not know of any multinational which has used personal data unlawfully and then been brought down by the European GDPR. However, the GDPR does make it difficult for researchers to gather personal data and use them to move forward with solutions to problems in society and people's health."

115 Petia Radeva (http://www.ub.edu/cvub/petiaradeva/?p=36).

“We could use people’s movement patterns to streamline transport in cities and promote more active and less sedentary lifestyles, which would save public healthcare money because it would prevent certain diseases, or more meticulous prevention work could be done. But we cannot do this using data from the entire population because researchers do not have access to this information. I would like to take part in a campaign to promote data donation like the ones run for organs or blood. Data donation should be promoted with transparency and explanations while ensuring that they are put to good use. In the United Kingdom there is a free database researchers can use. This is a welcome initiative to make further progress in science.”

“When an agency or a university asks you for data (for example, about your physical activity over the last month or year), your initial reaction is often fear or refusal. We don’t think they can be used to make progress in diseases which will save the lives of our children or grandchildren. But afterwards we happily hand them over to any company on the internet in exchange for entertainment. Why?”

Beyond technology

“An example: how is it possible that Spain, the standard-setter for the Mediterranean diet, has the highest rate of child obesity in Europe? We have a very high rate of child obesity because we do not have data about eating habits and we cannot make any progress. It would be as easy as families and schools gathering and providing them via mobile phones to make headway against this disease. Technology has taken a big step forward in data collection and processing capabilities. Now society has to do so as well.”

Distinguished Research Professor in the Department of Information and Communication Technology at Pompeu Fabra University (UPF). He leads the Web Science and Social Computing research group

CARLOS CASTILLO:[116] "It's very difficult for people to know that an algorithm is biased"

The Web Science and Social Computing Research Group at the UPF teamed up with the Technical University of Berlin and the Eurecat Technology Centre to build an algorithm which identifies and mitigates biases in other algorithms. They called the system FA*IR[117] and it works on other automated systems (for searching or recommending) which do not factor in the representation of different groups.

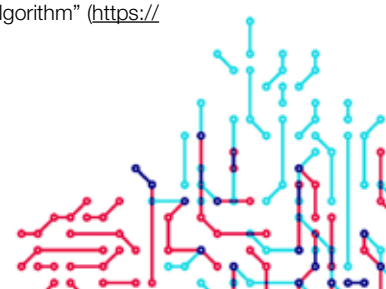
"For example, a person using this system may determine that there should be at least 20% or 40% women in the results, or people under 25, or people from such-and-such a background, or professionals from such-and-such sector," explains Castillo. "Biases are always there and they can be identified and mitigated."

Of course, if algorithms are already being used in many sectors, there will need to be lots of people identifying and mitigating the biases. How can this be done? "You need transparency," Castillo replies. "And the ability to hold the system to account and make it explain its decisions. However, as an individual you have very few tools to learn whether there is a powerfully biased algorithm behind it that is discriminating against you. If you search LinkedIn, journalists in Barcelona, you will most likely not be listed and there is most likely a gender bias based on past data. Until someone does a major study, like us in this case, and the results are known, it is impossible."

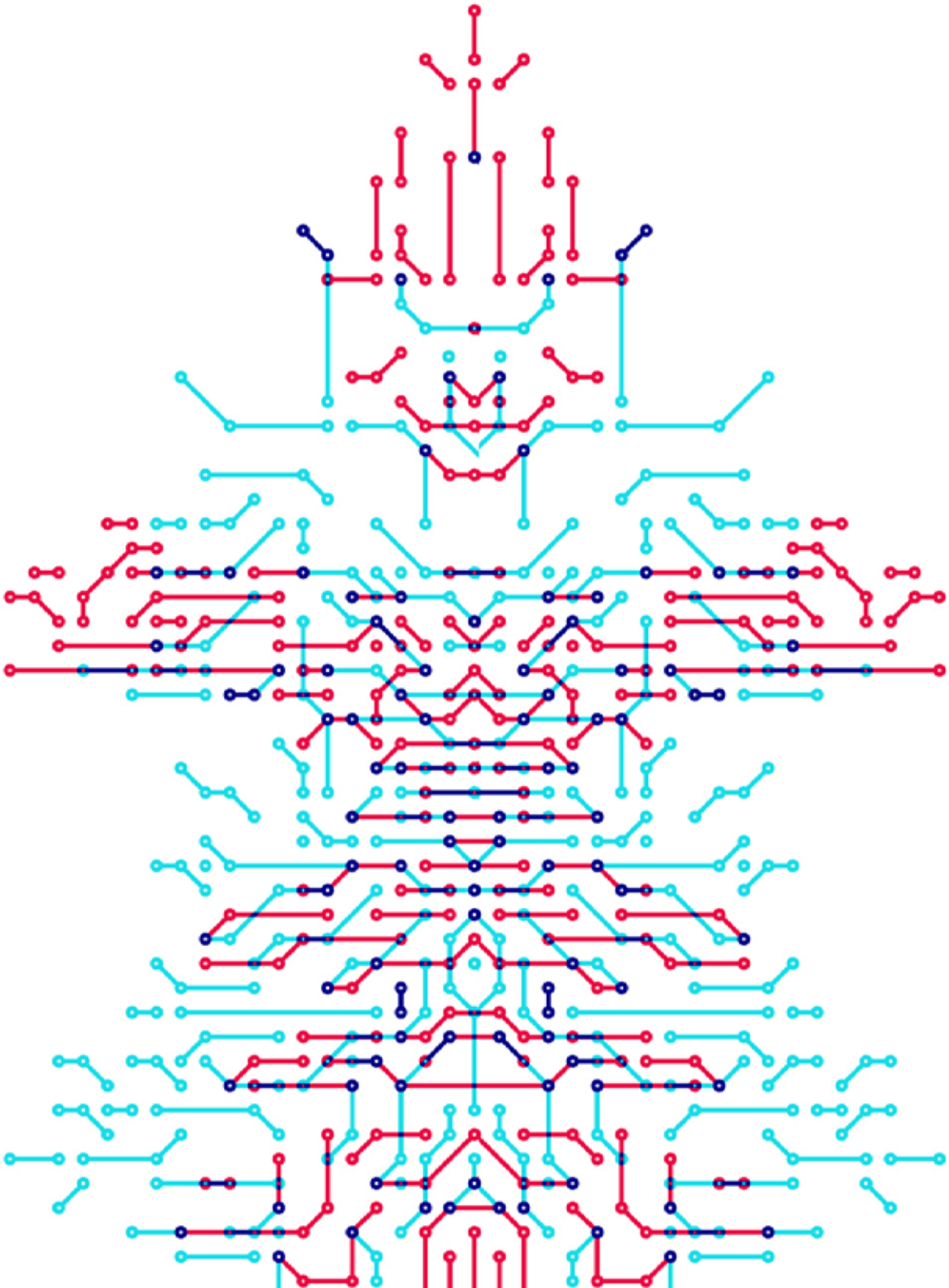
Algorithmic justice

"Some Catalan public agency ought to think about it. Perhaps the Catalan Data Protection Authority or a similar organisation could provide a solution so that citizens, consumers, have more information about the algorithms which make decisions. In France, for example, the National Commission on Information Technology and Liberties (CNIL)[118] is involved in these issues. The European *General Data Protection Regulation,* the GDPR,[119] also allows complaints by citizens' groups or associations."

116 Carlos Castillo (http://chato.cl/research/).
117 M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, R. Baeza-Yates. "FA*IR: A Fair Top-k Ranking Algorithm" (https://arxiv.org/pdf/1706.06368.pdf).
118 CNIL (https://www.cnil.fr/).
119 GDPR (https://gdpr-info.eu/).



“The European General Data Protection Regulation additionally enforces fair data processing; therefore, this fairness should be enforced in the algorithms,” he adds. “The GDPR allows citizen associations to file complaints. For example, if you think that women journalists are discriminated against on LinkedIn, you could gather data from lots of journalists and report it.”



Professor in the Department of Political and Social Science at Pompeu Fabra University (UPF)

CARLES RAMIÓ:[120] "Intensive use of AI and robotics is the only way to safeguard the welfare state"

Carles Ramió says that government is already behind the curve in implementing artificial intelligence (AI). This summer he published a 24-point code of ethics,[121] a draft of what we should already be thinking about. One of the most controversial points is to opt for a system of validation of the algorithms used by the public sector. "The process of designing and training the automated system has to be controlled," he argues. "It can't just be run by engineers. We need philosophers, sociologists, people who think about ethical issues, who take all kinds of discrimination into account. And if we buy algorithms from private firms, we have to disclose them as well."

Author of the book *La intel·ligència artificial i l'administració pública,*[122] he thinks that "there shouldn't be any black boxes. I'm not saying that everyone has to know the algorithm's formula. But it should be possible for a sentencing system to be reviewed, at least by experts and lawyers, to know what the training process was or why it wasn't taught more diversity."

"The government should put in place systems to certify algorithms to ensure they are diverse. It should have an ethical and technical vision team looking at the biases which would make the algorithm and the training system transparent."

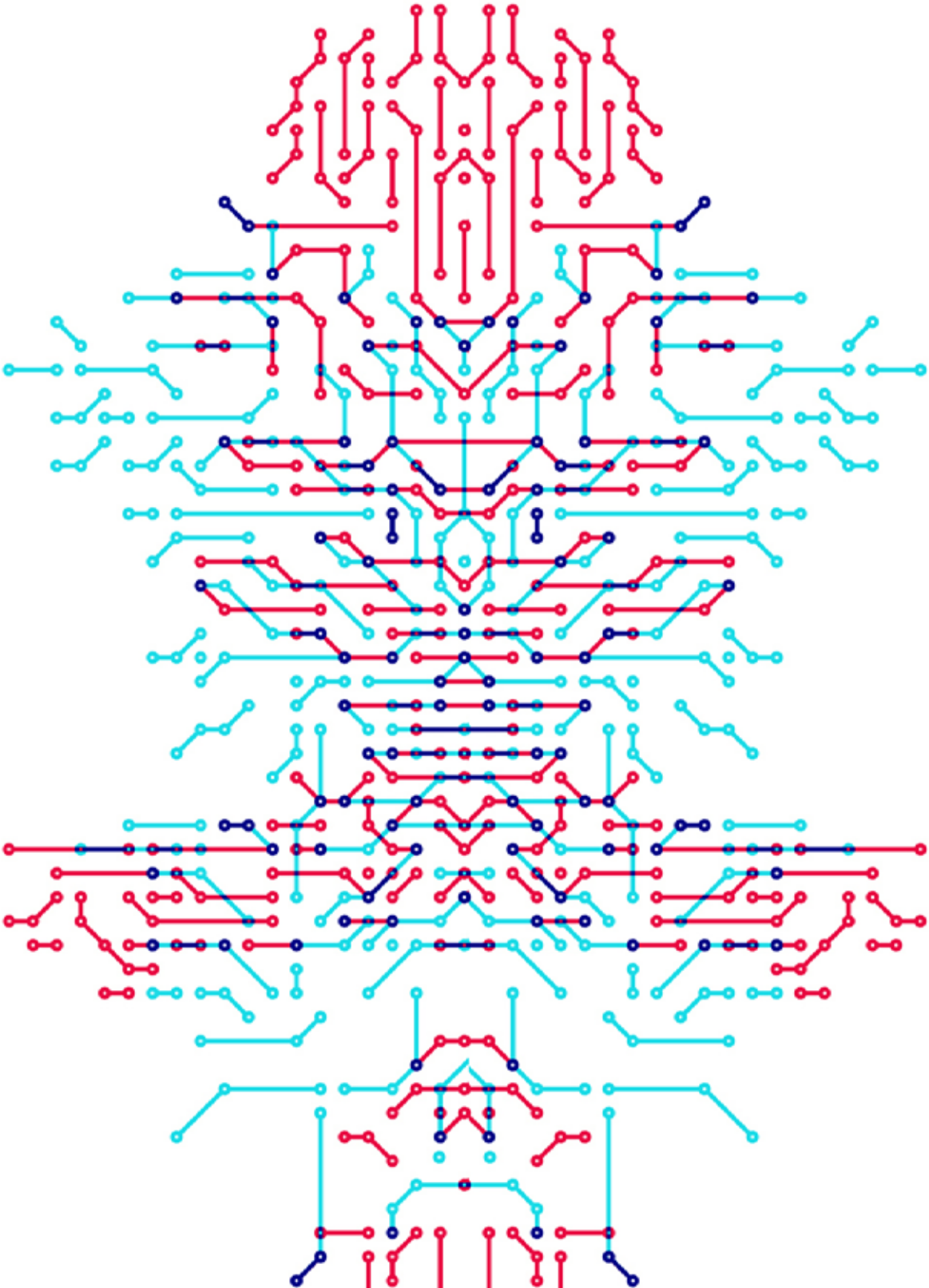
A great opportunity

"I see a huge opportunity in AI. It has to be rolled out as soon as possible. Well designed algorithms are a must. If all the data in an intensive care unit are monitored and you feed the algorithm properly, the system is not going to fail. It will detect any anomaly or disease. After that there is the scrutiny of the doctor who will assess and confirm the algorithm's diagnosis or not."

"Admittedly this involves an upfront investment. But the only way we still have something of a welfare state is due to the intensive use by government of AI and robotics. The ageing of the population means we cannot meet all the costs involved in health

120 Carles Ramió (https://www.upf.edu/es/web/politiques/entry/-/-/1182/401/carles-ramio).
121 Carles Ramió. *Estatuto ético para la implantación de la inteligencia artificial y la robótica en la Administración pública* (https://www.administracionpublica.com/estatuto-etico-para-la-implantacion-de-la-inteligencia-artificial-y-la-robotica-en-la-administracion-publica/).
122 *La intel·ligència artificial i l'Administració pública*. Los Libros de la Catarata (https://www.todostuslibros.com/libros/inteligencia-artificial-y-administracion-publica_978-84-9097-590-9).

and social services. Well-planned and targeted AI is achievable. But we need to change the discourse and be more proactive in government. The red flag is that algorithms are coming into government from private companies. But then how can we be transparent if we are reliant on the private sector?”



Distinguished Professor of Computer Science at Rovira i Virgili University (URV) in Tarragona. Researcher at the URV's ICREA-Acadèmia. Director of the UNESCO Chair of Data Privacy. Director of CYBERCAT - Centre for Research in Cybersecurity of Catalonia

JOSEP DOMINGO:[123] "Intelligence means 'understanding' and the machine does not understand"

Like the other experts interviewed, Domingo thinks we have a serious problem with artificial intelligence and most of all with deep learning. That's because although they are tremendously efficient, there is no way of knowing how the machine has learned. "The moment a person applies for a loan on a website and after they have entered some data the automated system says their application has been rejected, they will want an explanation. And here comes the problem."

"We have less and less knowledge of why things happen. The stockbroker is dying out as it is very hard to compete with a robot investor who takes economic variables from all over the world and in a split second decides what to do. Algorithmic explainability[124] or transparency is an attempt to retrieve knowledge. You cannot fall back solely on the machine. You have to understand what the world is about; otherwise you lose control."

The right to an explanation

"The machine is not intelligent. Nor does it know why it has decided what it has decided. Intelligence means 'understanding' and the machine does not understand. It is trained with some historical data (e.g. data from past loans which applicants paid back or didn't pay back) to fine-tune the parameters of the decision-making algorithm. This is where the learning process ends. Afterwards, it is given a specific case about which a decision has to be made (e.g. a new loan application) and the machine-algorithm calculates the decision using the learned parameters. Sticking to the loan example, if for whatever reason the algorithm finds that the new applicant resembles people who did not pay back their loan in the past, the decision will be to reject the application, and vice versa.

"The European data protection regulation (the GDPR) enshrines the right to an explanation which is basically about protecting democracy. Article 22 is the touchstone. With this legal requirement, we realise that there is no way of providing satisfactory explanations. Every time someone challenges the decision of a machine, engineers have to try

123 Josep Domingo Ferrer (http://www.urv.cat/es/universidad/conocer/personas/profesorado-destacado/2/josep-domingo-ferrer).
124 "Algorithmic explainability or transparency" (https://en.wikipedia.org/wiki/Explainable_artificial_intelligence).

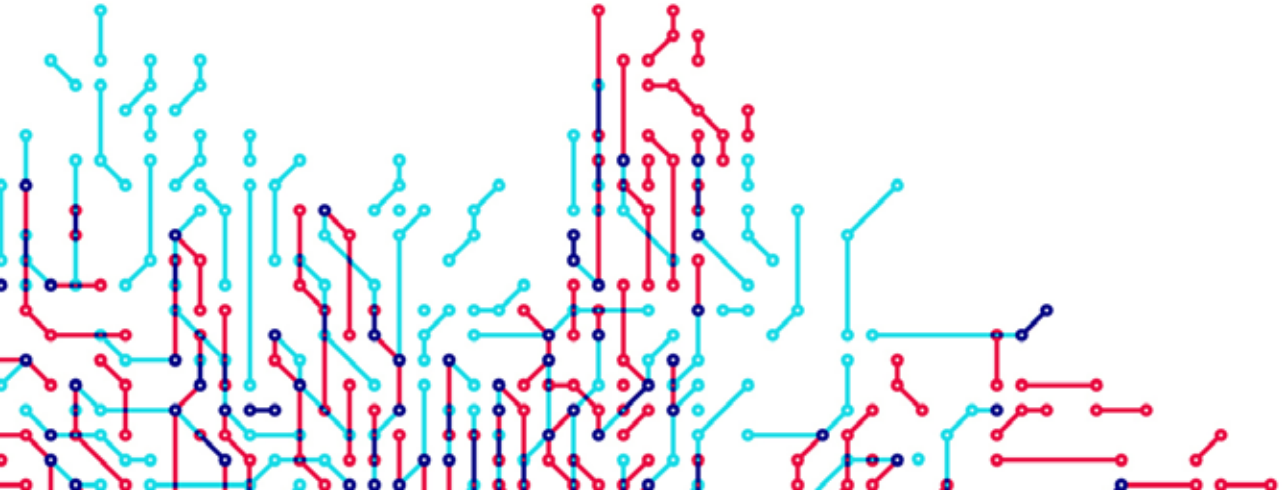
to reconstruct what it has done by hand. But that is not scalable. We don't have the tools to give automatic explanations."

"Fortunately for anyone who uses automated decision-making algorithms, the public still does not fully understand what is going on. The same thing happened with data protection. But I haven't seen any demonstrations in the streets, not even with the Cambridge Analytica and Facebook case,[125] about breach of privacy."

Accountability for the algorithm

Black boxes are a concern and therefore algorithmic transparency is called for. But there are also concerns about specifying who is accountable for the algorithm's decision. When autonomous cars are driving around without drivers and someone gets run over, who will be accountable? The analyst who designed it? The programmer? The manufacturer of the car for selling it?

Domingo provides another perspective: "The car manufacturer, the banker, the doctor or the politician commission the algorithm. And they have a rough idea of what the automated system should do. The analyst draws up a design and passes it on to the computer programmer. In between there's a lot of sketchy instructions. Finally, the algorithm should be checked again by the analyst and by the person who commissioned it, but this seldom happens. Moreover, if the training data have not been carefully chosen, it may be that the algorithm's learning is only valid for white men and Christians and that the special features of women, people of colour or Muslims, for example, have been neglected."



125 Karma Peiró. "L'escàndol Facebook-Cambridge Analytica: un cas per revisar la protecció de dades i molt més" (https://www.naciodigital.cat/noticia/151744/escandol/facebook-cambridge/analytica/cas/revisar/proteccio/dades/molt/mes).

Researcher in the Knowledge Engineering and Machine Learning group in the Intelligent Data Science and Artificial Intelligence Research Centre at the UPC. Vice-President for Big Data, Data Science and Artificial Intelligence at the Computer Engineering Association of Catalonia

KARINA GIBERT:[126] "The energy cost of AI is huge"

"In principle technology is neutral and what is more or less ethical is what we do with it," says researcher Karina Gibert. "The ethical problems to be borne in mind today are biases but also explainability, which is a much deeper issue and has already been discussed in this section."

Gibert points out that there is only one part of AI which cannot be explained and which is known as subsymbolic. "The symbolic part which imitates the way humans solve problems is totally explainable. But in computational terms it's horribly expensive and has never really made the leap to real problems of some complexity because it gets jammed up."

"The subsymbolic part seeks to get computers to resolve problems which call for intelligence with greater quality than humans from the standpoint of the results it provides and regardless of whether or not the way it resolves the problem mimics the way humans do it," she says. "And this is the unexplainable AI which among other things has led to deep learning. In return, deep learning has succeeded in delivering results for major problems which we hadn't been able to solve until now. However, if the outcome of an artificial intelligence process is to have any impact on a real life situation, obviously the decision which has been taken will have to be justified in some way. And those of us who like me have been working on real applications for thirty years are fully aware of this."

AI's ecological footprint

Gibert also adds a new insight into ethics which has not been mentioned thus far. "The cost of the power required to run all these algorithms and to host all the data involved in these processes is enormous. And it has an ecological footprint. The question is: do we need your watch to measure body temperature every five minutes over a year when we know that the temperature changes pretty uniformly and there are visible changes every hour?"

"This is closely related to a dilemma that researcher Ricardo Baeza-Yates often mentions: 'big data or right data?' Perhaps we need to have fewer data and they should

126 Karina Gibert (https://www.eio.upc.edu/en/homepages/karina).

be informative and representative, which would save us the energy we use store and process them. In terms of sustainability it is crucial to consider what kind of algorithms and data usage we have."

Who is accountable if the algorithm fails?

"I think there are two scenarios. 1. Non-adaptive artificial intelligence (it does not change the operating mechanisms and retains the design), where the responsibility lies with the designer of the algorithm. I think it's pretty obvious that whoever designs the algorithm should be fully aware about the risk values associated with the decisions or recommendations the system may make. But I also think that perhaps the levels of uncertainty associated with the algorithm's solutions are not made sufficiently apparent. This is like the guidelines in medicine which say: 'When a person comes in with a heart attack, this is the dose you should give them.' There is also a lot of uncertainty about the standard dose and it may not be appropriate for a person in particular. And when this happens the doctor is not responsible in any way if they have followed the guideline. I don't really see much difference in how what artificial intelligence recommends should be dealt with as opposed to what a doctor decides, as it is also based on standard recommendations which may not work in a particular case."

"Scenario 2. When AI is adaptive. It is hard to know what kinds of actions it might take in the long term. If artificial intelligence keeps a record of what it does and alters its behaviour depending on what works better or worse based on previous actions, we may find that after a certain period of time reasoning or recommendations emerge which we could not have envisaged. In other words, AI would be going further on its own and creating new solutions. Here it is really difficult to require or believe that whoever designed the algorithm could have checked out all the situations and predict and restrict ones which might be hazardous for humans. This sets up a debate about the extent to which we want artificial intelligence to express itself to its full potential when dealing with people or environments in which it interacts with humans and may pose risks."

Gibert suggests artificial intelligence should be audited (just as associations audit software in general) in case of malfunction to find out whether it is because the problem has been poorly described or due to poor design, poor implementation or system degeneration. In each case the responsibility will be borne by one person or another. "However, it is up to the developer to ensure whoever commissions the algorithm or its role thinks about all the limits, risks and problems so that the design is as complete as possible and its use as appropriate as possible. This is not done enough and should be done more often."

2.6. Unresolved dilemmas

Scientist Marie Curie said at some point in her long career: “Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.”

In the previous section, the leading artificial intelligence researchers in Catalonia exhaustively described the dilemmas we face today with AI such as biases in the algorithms, the lack of explainabilty when they operate with deep learning and the absence of consensus on who should be accountable if the machine makes a mistake.

Other questions will crop up in the future which also have no clear answer today. “How can we ensure that automated decision-making (or the actions that flow from it) do not have a harmful impact on people? What levels of security do these smart systems have to ensure that they are not vulnerable to cyberattacks or misuse? What will happen when an algorithm knows us better than we do ourselves and can manipulate our behaviour subliminally?” asked Núria Oliver, a PhD in Artificial Intelligence from the Massachusetts Institute of Technology (MIT), in the inaugural lecture of the 2019-2020 academic year of the Catalan university system.[127]

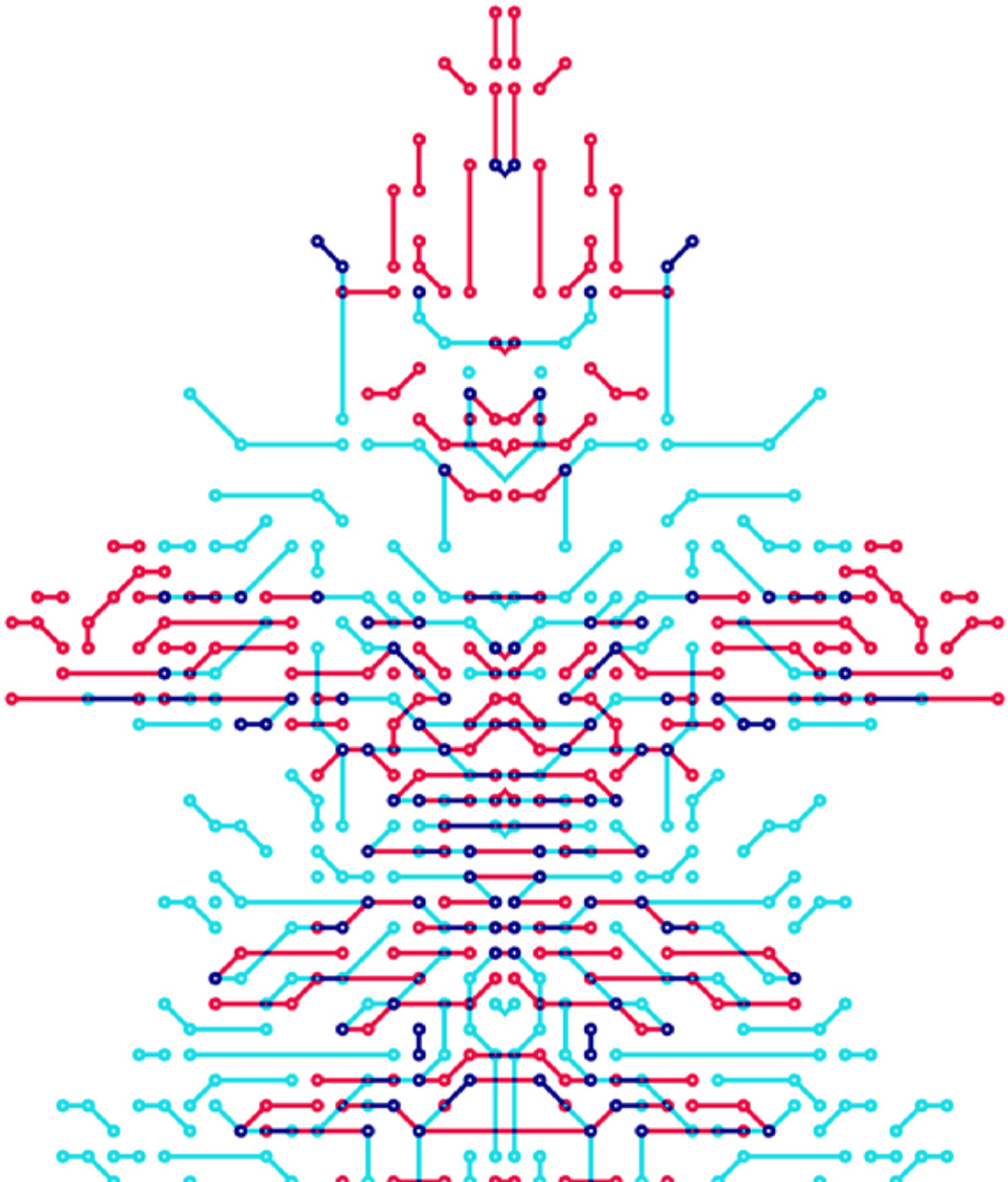
In lockstep with the advances in automated decision-making algorithms (ADAs), there are two issues which we are not paying enough attention to and are becoming increasingly important in the technological societies in which we live: trust and transparency.

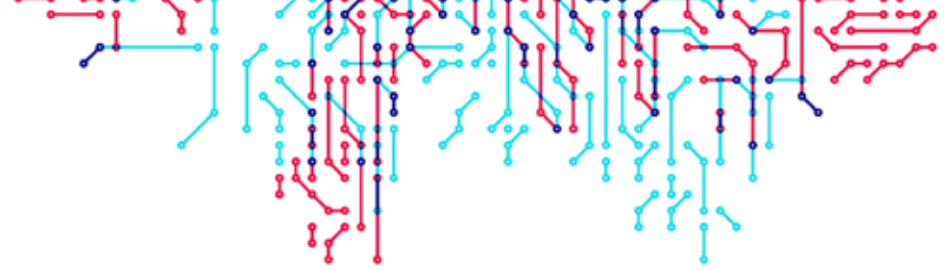
1. **Trust** is a basic pillar in the relationship between humans and institutions in any society. Technology needs the trust of users who are increasingly turning their lives over to digital services. Yet in recent years we have changed the meaning of the term. Through technology we trust strangers when we let a room in our house through Airbnb and open the door to people we do not know; when we are taken to a destination without knowing whether the person at the wheel drives well or badly, be it BlaBlaCar or via Uber; when we transfer money using a mobile app from a company we have no prior experience of. But trust has slipped through our fingers, especially after scandals such as Cambridge Analytica which illegally handled 50 million people’s data and undermined the meaning of the word *democracy*.
2. And **transparency**… A term widely used in recent years but very inconsistent because what we really have is technological murkiness. “A computer system is transparent when a non-expert person looks at it, understands it and knows how it works,” says Núria Oliver. “But this is not the case today.” She says that there is increasing opacity, whether because businesses want to protect the intellectual property of their algorithms, or because the public does not have the core technological

127 Nuria Oliver’s lecture at the opening of the 2019-2020 academic year in the Catalan university system (https://www.youtube.com/watch?v=DwC0KDwliXc).

knowledge needed to understand the explanations, or because governments do not give transparency the importance it deserves (although the word is in almost all party platforms), or because deep learning prevents us from getting an explanation of why it has taken a particular decision.

AI is growing rapidly worldwide. Whoever masters it will not only have economic power but also political and social power. At this point the only approach that we as citizens should consider is to achieve (and demand) trustworthy artificial intelligence, designed and thought out for people, with strong ethical awareness and which conforms to the values of justice, (genuine) transparency and fairness. Secure artificial intelligence which can be audited and where we can ask for explanations about it.





2.7. Recommended reading for further insights

This section includes public standards regulating the use of artificial intelligence in Europe and also publications recommended by the abovementioned researchers which have provided context and documentation for this research.

Strategies, declarations and recommendations

Artificial Intelligence Strategy of Catalonia (https://participa.gencat.cat/processes/estrategiaIA).

"Barcelona Declaration for the proper development and usage of artificial intelligence in Europe" 2017 (https://www.iiia.csic.es/barcelonadeclaration/).

"Declaration on Ethics and Data Protection in Artificial Intelligence" (https://apdcat.gencat.cat/web/.content/04-actualitat/noticies/documents/ICDPPC-40th_AI-Declaration_ADOPTED.pdf).

Ethical guidelines for trustworthy AI (https://ec.europa.eu/spain/barcelona/news/press_releases/190408_ca).

Spanish R&D strategy in artificial intelligence (http://www.ciencia.gob.es/stfls/MICINN/Ciencia/Ficheros/Estrategia_Inteligencia_Artificial_IDI.pdf)

OECD recommendations on artificial intelligence (https://www.oecd.org/centrodemexico/medios/cuarentaydospaisesadoptanlosprincipiosdelaocdesobreinteligenciaartificial.htm).

European Parliament Resolution "On a comprehensive European industrial policy on artificial intelligence and robotics". February 2019 (https://www.europarl.europa.eu/doceo/document/A-8-2019-0019_EN.html).

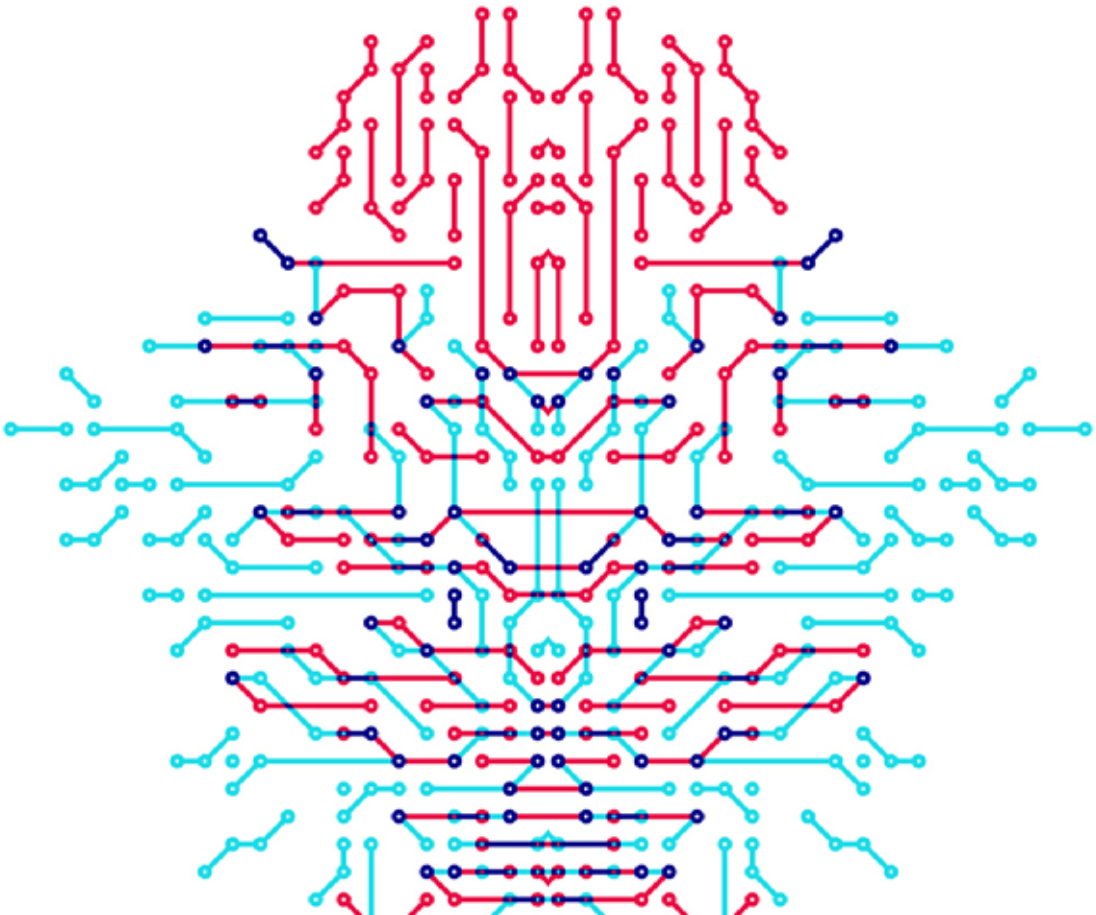
EU artificial intelligence ethics checklist ready for testing as new policy recommendations are published (https://ec.europa.eu/digital-single-market/en/news/eu-artificial-intelligence-ethics-checklist-ready-testing-new-policy-recommendations-are).

Regulating big data. The guidelines of the Council of Europe in the context of the European data protection framework (http://isiarticles.com/bundles/Article/pre/pdf/106203.pdf).

Artificial Intelligence and Data Protection: Challenges and Possible Remedies (https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6).

Principles established by the European Commission's High-Level Expert Group on Artificial Intelligence (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

Alessandro Mantelero. *Artificial Intelligence and Data Protection: Challenges and Possible Remedies* (https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6).



Reports

Algorithm Watch. *Automating society. Taking Stock of Automated Decision-Making in the EU* (https://algorithmwatch.org/en/automating-society/).

Artificial Intelligence in Education. Challenges and Opportunities for Sustainable Development. UNESCO (https://unesdoc.unesco.org/ark:/48223/pf0000366994).

World Health Organisation releases first guideline on digital health interventions. WHO (https://www.who.int/news-room/detail/17-04-2019-who-releases-first-guideline-on-digital-health-interventions).

La intel·ligència artificial a Catalunya. Acció. Generalitat de Catalunya (https://www.accio.gencat.cat/ca/serveis/banc-coneixement/cercador/BancConeixement/la-intelligencia-artificial-a-catalunya).

La ciberseguretat a Catalunya Acció. Generalitat de Catalunya. 2018 (https://www.accio.gencat.cat/web/.content/bancconeixement/documents/informes_sectorials/ciberseguretat-informe-tecnologic.pdf).

Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. IEEE. Advancing Technology for Humanity. 2016 (http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf).

Partnership on AI to benefit people and society. Website (https://www.partnershiponai.org/#).

European Civil Law Rules in Robotics. European Parliament. 2016 (http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf).

La Internet de les coses a Catalunya. Acció. Generalitat de Catalunya (https://www.accio.gencat.cat/web/.content/bancconeixement/documents/pindoles/iot-cat.pdf).

Al Pascual. *2015 Data Breach Fraud Impact Report* (https://www.javelinstrategy.com/coverage-area/2015-data-breach-fraud-impact-report).

Roy Wedge, James Max Kanter, Kalyan Veeramachaneni, Santiago Moral Rubio, Sergio Iglesias Perez. *Solving the false positives problem in fraud prediction using automated feature engineering* (http://www.ecmlpkdd2018.org/wp-content/uploads/2018/09/567.pdf).

Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.* AI Now Institute (https://ainowinstitute.org/aiareport2018.pdf).

AI in Context. Data & Society (https://datasociety.net/output/ai-in-context/).

The Cybersecurity Campaign Playbook. 2018 (https://www.ndi.org/publications/cybersecurity-campaign-playbook-global-edition).

Itziar de Lecuona. *Evaluación de los aspectos metodológicos, éticos, legales y sociales de proyectos de investigación en salud con datos masivos (*big data*)* (http://scielo.isciii.es/scielo.php?script=sci_abstract&pid=S0213-91112018000600576).

El vehicle connectat a Catalunya. Acció. Generalitat de Catalunya. 2019 (https://www.accio.gencat.cat/ca/serveis/banc-coneixement/cercador/BancConeixement/vehicle-connectat-a-catalunya).

Liliana Arroyo. *Trustful and trustworthy: Manufacturing trust in the digital era* (https://www.esade.edu/researchyearbook/node/231).

Eli Pariser. *The Filter Bubble: What the Internet Is Hiding From You* (https://hci.stanford.edu/courses/cs047n/readings/The_Filter_Bubble.pdf).

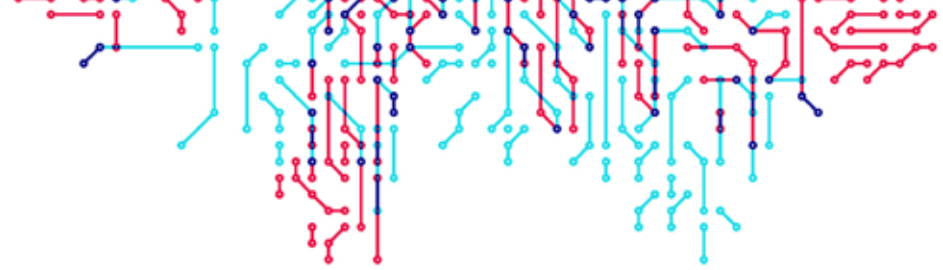
Articles

Itziar de Lecuona. “La oportunidad de la investigación con datos masivos en salud” (http://www.gacetasanitaria.org/es-evaluacion-los-aspectos-metodologicos-eticos-articulo-S0213911118300864).

Ramón López de Mántaras. “La ética en la inteligencia artificial” (https://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/el-multiverso-cuntico-711/tica-en-la-inteligencia-artificial-15492).

Ricardo Baeza-Yates, KarmaPeiró. “És possible acabar amb els biaixos dels algorismes? (1a i 2a part)”. 2019 (https://www.karmapeiro.com/2019/06/17/es-possible-acabar-amb-els-biaixos-dels-algoritmes-1a-part/).

Tomer Hochma. “The Ultimate List of Cognitive Biases: Why Humans Make Irrational Decisions” (https://humanhow.com/en/list-of-cognitive-biases-with-examples/).



"Machine Bias". *ProPublica* (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Sandra Wachter. "How to make algorithms fair when you don't know what they're doing". *Wired*. (https://www.wired.co.uk/article/ai-bias-black-box-sandra-wachter).

Jayshree Pandya. "Hacking Our Identity: The Emerging Threats from Biometric Technology" (https://www.forbes.com/sites/cognitiveworld/2019/03/09/hacking-our-identity-the-emerging-threats-from-biometric-technology/#353ed3505682).

Loren Grush. "Google apologizes after Photos app tags two black people as gorillas" (https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas).

Kyle Wiggers. "MIT researchers: Amazon's Rekognition shows gender and ethnic bias (updated)" (https://venturebeat.com/2019/01/24/amazon-rekognition-bias-mit/).

Karma Peiró, Ricardo Baeza-Yates. "Algoritmo, yo también existo" (https://www.karma-peiro.com/2019/09/24/algoritmo-yo-tambien-existo/).

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, Ricardo Baeza-Yates. "FA*IR: A Fair Top-k Ranking Algorithm" (https://arxiv.org/pdf/1706.06368.pdf).

Carles Ramió. "Estatuto ético para la implantación de la inteligencia artificial y la robótica en la administración pública" (https://www.administracionpublica.com/estatuto-etico-para-la-implantacion-de-la-inteligencia-artificial-y-la-robotica-en-la-administracion-publica/).

La intel·ligència artificial i l'administració pública. Los Libros de la Catarata (https://www.todostuslibros.com/libros/inteligencia-artificial-y-administracion-publica_978-84-9097-590-9).

Karma Peiró. "L'escàndol Facebook-Cambridge Analytica: un cas per revisar la protecció de dades i molt més" (https://www.naciodigital.cat/noticia/151744/escandol/facebook-cambridge/analytica/cas/revisar/proteccio/dades/molt/mes).

Agustí Cerrillo. "Com obrir les caixes negres de les administracions públiques? Transparència i rendició de comptes en l'ús dels algoritmes" (http://revistes.eapc.gencat.cat/index.php/rcdp/article/view/10.2436-rcdp.i58.2019.3277).

Videos

Nuria Oliver's inaugural lecture for the 2019-2020 academic year in the Catalan university system (https://www.youtube.com/watch?v=DwC0KDwliXc).

Por la razón y la ciencia. Chilean television documentary. Intervention by Ricardo Baeza-Yates (https://www.youtube.com/watch?time_continue=1253&v=7PCC7tRyM2I).

2.8. Acknowledgements

Finally, I would like to add my sincere thanks to the experts in artificial intelligence and data science who have been interviewed and who have enabled us to grasp the current state of the art in the use of automated decision-making algorithms in Catalonia in terms of both risks and rewards. My thanks also go to the heads of government ministries and leaders of businesses who have made it possible to share more than fifty examples.

The idea was to explore this technology in depth while avoiding the usual scaremongering of out-of-context headlines. The personalised interviews and the time afforded by all the people mentioned below are priceless.

Ricardo Baeza-Yates, CTO at NCENT. Professor of Computer Science at Pompeu Fabra University and Northeastern University.

Anton Bardera, lecturer in the Department of Computer Science, Applied Mathematics and Statistics at the University of Girona (UdG).

Meritxell Bassolas, Director of Knowledge and Technology Transfer at the Computer Vision Centre at the Autonomous University of Barcelona (UAB).

Marga Bonmatí, Director of the AOC Consortium.

Marco Bressan, former lead data scientist at BBVA. Director of Satellogic.

Victòria Camps, Professor of Ethics and Philosophy of Moral and Political Law at the Autonomous University of Barcelona (UAB).

Carlos Castillo, Distinguished Research Professor in the Department of Information and Communication Technology at Pompeu Fabra University (UPF). He leads the Web Science and Social Computing research group.

Ulises Cortés, Scientific Director of the High Performance Artificial Intelligence group at the Barcelona Supercomputing Centre. Professor of Artificial Intelligence at the Technical University of Catalonia (UPC).

Fernando Cucchietti, Director of Data Analysis and Visualisation at the Barcelona Supercomputing Centre.

Josep Domingo, Distinguished Professor of Computer Science at Rovira i Virgili University (URV) in Tarragona. Researcher at the URV's ICREA-Acadèmia. Director of the UNESCO Chair of Data Privacy. Director of CYBERCAT - Centre for Research in Cybersecurity of Catalonia.

Ricard Gavaldà, lecturer and coordinator of the research laboratory at the Technical University of Catalonia (UPC). CEO and scientific director at Amalfi Analytics.

Karina Gibert, researcher in the Knowledge Engineering and Machine Learning group in the Intelligent Data Science and Artificial Intelligence Research Centre at the Technical University of Catalonia (UPC). Vice-President for Big Data, Data Science and Artificial Intelligence at the Computer Engineering Association of Catalonia.

Elisabet Golobardes, PhD in Computer Engineering and Professor of Artificial Intelligence and Data Science at La Salle-Ramon Llull University. Member of the Date Science for the Digital Society (DS4DS) research group.

Àlex Hinojo, former Director General of the Amical Wikimedia association. Co-founder of the Drets Digitals project.

Alberto Labarga, data scientist at IOMED.

Itziar de Lecuona, lecturer in the Department of Medicine at the University of Barcelona (UB). Deputy Director of the Bioethics and Law Observatory at the UB.

David Llorente, CEO of Narrativa.

Jorge López, Head of the Research and Development Unit at UBinding.

Ramon López de Mántaras, Research Professor at the Spanish National Research Council (CSIC). Former director of the Artificial Intelligence Research Institute (IIIA).

Gabriel Maeztu, doctor and data scientist. CEO at IOMED.

Alessandro Mantelero, Council of Europe rapporteur on Artificial Intelligence and Data Protection. Professor of Law at the University of Turin.

Jordi Mas, member of Softcatalà and a pioneer of the Catalan Internet.

Manel Medina, professor at the Technical University of Catalonia (UPC). Founder and Director of esCERT-UPC.

Jordi Navarro, CEO & co-founder of Cleverdata.io.

Adina Nedelea, Head of the Mathematics Team at UBinding.

Ivan Ostrowicz, engineer and expert in artificial intelligence applied to education.

Petia Radeva, lecturer in the Department of Mathematics and Computer Science at the University of Barcelona (UB). Director of the Machine Learning and Computer Vision consolidated research group at the UB.

Carles Ramió, professor in the Department of Political and Social Science at Pompeu Fabra University (UPF).

Teresa Roig Sitjar, technology and science adviser.

Pier Paolo Rossi, Director of Advanced Customer Marketing & Analytics at Banc de Sabadell.

Giusseppe Scionti, CEO and founder of NovaMeat.

Carles Sierra, Director of the Institute for Research in Artificial Intelligence (IIIA).

Antonio Andrés Pueyo, senior researcher in the Group of Advanced Studies on Violence (GEAV) at the University of Barcelona (UB). Professor of Psychology at the UB.

Carme Torras, PhD in Computer Science and Research Professor at the Institute of Robotics and Industrial Computing (CSIC-UPC).

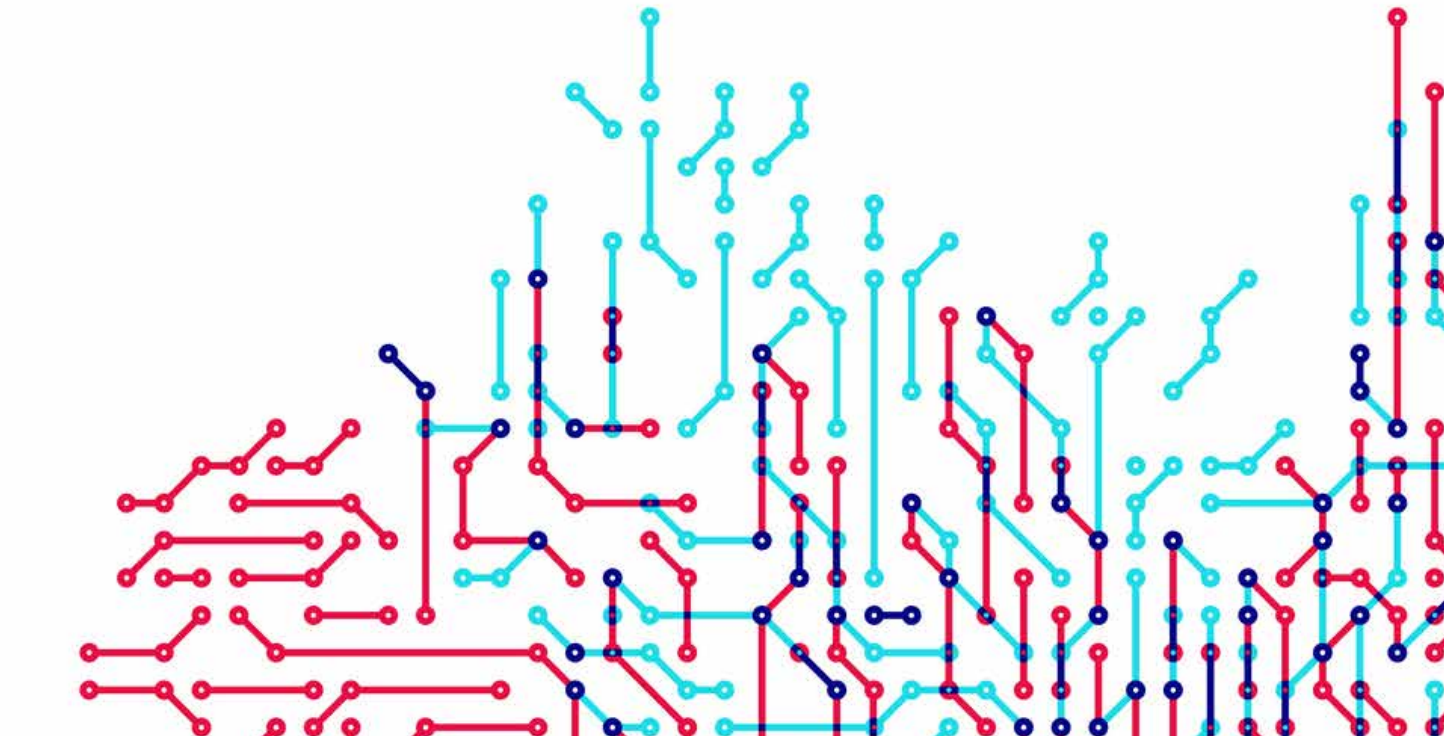
Jordi Torras, founder and CEO of Inbenta.

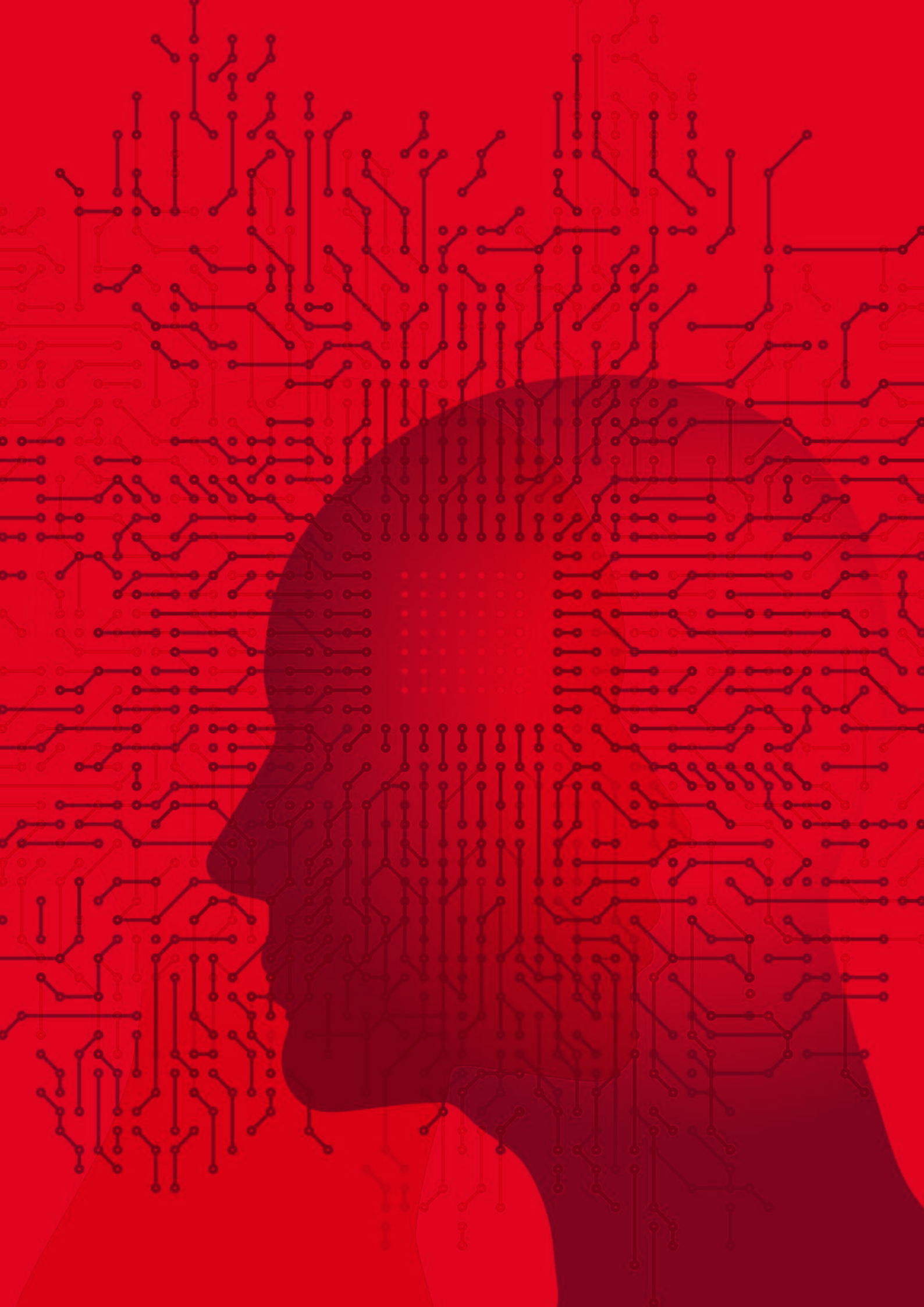
Lluís Torrens, Director of Social Innovation in Barcelona City Council's Social Rights, Global Justice, Feminism and LGTBI Department.

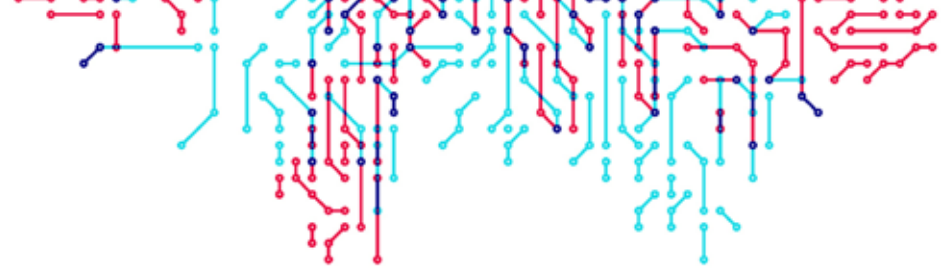
Jordi Vitrià, Professor of Computer Languages and Systems at the University of Barcelona (UB). Member of the Department of Mathematics and Computer Science at the UB.

Staff in the **Technical Secretariat at the** Catalan Public Employment Service.

This section is the result of the fieldwork conducted by Karma Peiró, an ICT journalist.







3.
DATA PROTECTION AND AI

In the introduction, when discussing the risks associated with automated decision-making algorithms (ADAs), we emphasised two types of risks: those related to their specific use (e.g. the risk of being denied a loan) and those related only to the processing of personal data (e.g. the risk of one's data being leaked). The fieldwork on ADA uses in Catalonia has focused on the risks associated with each of the uses presented. In this section of the report, we will focus on the risks associated with data protection.

Respect for data protection is essential in order to ensure the ethical use of automatic decision-making algorithms. Within the context of the European Union, this means compliance with the GDPR.

When talking about the risks associated with data protection, we mean the failure to comply with one or more provisions of the GDPR. The intensive use of personal data involved in many practical applications of ADAs leads to conflict between ADAs and the GDPR. In this scenario, the objective of the Catalan Data Protection Authority and other supervisory authorities is to ensure compliance with the GDPR's provisions.

The illegitimate collection and processing of data

The GDPR requires all data processing to have a legitimising legal basis. Several exist, and they all are equally valid: consent, legal obligation, legitimate interest, etc. Irrespective of the legal basis used, data subjects must be informed of the processing of their data.

Nowadays, personal data are very valuable. It has been said that data are the new oil. In this context, it would be naive to think that all organisations that collect personal data do so in an appropriate manner. The high levels of personalisation allowed by ADAs provide an incentive to collect as much personal data as possible since they are the basis for their operation.

Many examples of illegitimate data collection and processing exist. Focusing particularly on mobile devices (which have enormous potential for generating personal data because people spend so much time each day in close contact with them), there are two very interesting research papers that refer to the illegitimate use of data. The first[128] analyses two techniques used by some mobile apps available in app stores to breach Android

128 Joel Reardon, Álvaro Feal, *et al*. "50 ways to leak your data: an exploration of apps' circumvention of the Android permission system". *28th USENIX Security Symposium*, 2019.

controls in order to gain access to data. The second[129] explains how software programs installed during a mobile's manufacturing process take advantage of their privileges to read and distribute data without users' knowledge.

The right not to be subject to automated decision-making

ADAs make decisions based on people's profiles (characteristics of a person that the algorithm's designers have deemed to be relevant). Limiting decision-making to a set of previously defined characteristics is a major restriction on a human judge's ability to assess other aspects that may on the face of it seem to be largely irrelevant. For example, by talking to a human interlocutor a person may explain the importance or irrelevance of some of the characteristics of their profile or even point out things that the profile does not reflect.

The GDPR recognises this problem and establishes the right not to be subject to automated decision-making if such decisions may affect someone significantly on a personal level. In particular, Article 22 states:

"The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly affects him or her."

Article 22 does not prohibit the use of automatic decision-making but does require human intervention when a decision will have significant effects on people.

Although the wording of Article 22 mentions the right not to be subject to an automated decision, the guidelines issued for its interpretation state that it is not a right which data subjects have to demand but rather a prohibition. The distinction is significant because if this were a right of objection, organisations could habitually use automated decision-making and change their behaviour only in specific cases when a data subject requests it.

The principle of transparency

In order for individuals to be able to exercise the rights they have over their data, processing must be carried out in a transparent way. According to the GDPR, transparency requires all information provided to individuals (the data being processed, the purpose of the processing, the rights of individuals, etc.) to be concise, easily accessible and expressed in understandable language.

129 *Julien* Gamba*, Mohammed Rashed, et al. "An analysis of pre-installed Android software". 41st IEEE Symposium on Security and Privacy, 2020*

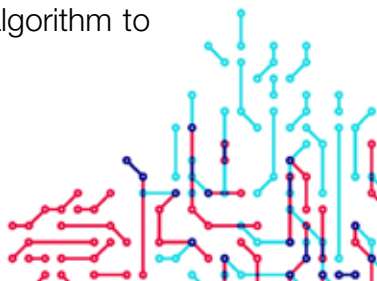
As regards the transparency of decisions taken by algorithms, the provisions of the GDPR are limited. While the recitals (the non-operative sections of the regulation that aid in its interpretation) mention the right to obtain an explanation of decisions, the regulation's operative section merely requires data subjects to be given relevant information on the logical basis for a decision. While providing an explanation for a decision requires dealing with the specific case, giving relevant information about the reasons for it may be done in a more general way with respect to the algorithm.

The decision not to ask for an explanation for individual decisions may be based on the technical inability to provide one in specific cases. Each AI algorithm has a specific level of explainability. In general, increases in the complexity of algorithms have been accompanied by reductions in the ability to explain their results. For example, while explaining a decision is simple in a rule-based system, it may be very complicated in a deep learning system (where a complex structure of artificial neurons self-regulates in the training phase of the model without external intervention).

Given the importance of machine learning, and in particular deep learning, this lack of explainability may be considered as one of artificial intelligence's weaknesses. But why is it important for an algorithm's results to be explainable? Sometimes we are not interested in knowing the reason for a decision; it is enough to know that the algorithm has a fairly good level of accuracy. For example, in the case of a film recommendation system, it may not seem particularly important to explain the reason for a specific recommendation. However, a specific explanation becomes more necessary when the decision may significantly impact people. Would you trust a doctor who does not give any explanation of their diagnosis, or a judge who issues rulings without justification? Hence if we demand explanations from human experts, why not ask ADAs to provide them as well?

The GDPR avoids this problem by requiring human intervention in decisions that may significantly affect people; that is, in cases of decisions for which an explanation is more necessary.

Furthermore, the need for some decisions to be explained is driving research in explainable artificial intelligence (XAI). In this regard, there are two approaches. The first seeks an explanation for an ADA's observed behaviour. This approach to an extent follows experimental science procedures: given the existence of an observed behaviour, we build a model that explains it. The problem is the validity of this model. As British statistician George Box said in 1978, all models that attempt to explain reality are wrong, but some are useful. Having a useful model to explain how an ADA works does not mean that the explanation for a particular case is the right one. The second approach avoids the problem of an explanation's validity by requiring the decision-making algorithm to



be easily interpretable. This is mainly achieved by reducing complexity. The price to be paid in this case is the loss of the ADA's accuracy.

Just as not all decisions are simple, it is also unnecessary to have an explanation for all decisions. The first step is to determine whether an ADA's decisions need to be explainable, and if they do to then decide how best to obtain the explanations. For example, in the case of an autonomous car, explanations of decisions may be useful for determining accountability in the event of an accident.

The principle of fairness

The principle of fairness requires data to be used in a way that is foreseeable by individuals (in relation to the intended purpose) and that does not inflict unjustified adverse consequences on individuals.

The main point of conflict between ADAs and the principle of fairness is the discrimination against individuals that may arise from an algorithm making biased decisions. ADAs must not treat a group of people unequally based on gender, skin colour, religious beliefs, sexual orientation, etc.

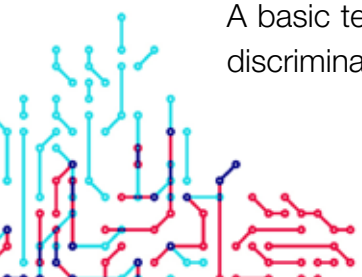
While we may believe that algorithms, unlike people, are not prejudiced and will therefore not make discriminatory decisions, this view is inconsistent with reality: ADAs have biases. The sources of these biases are diverse.

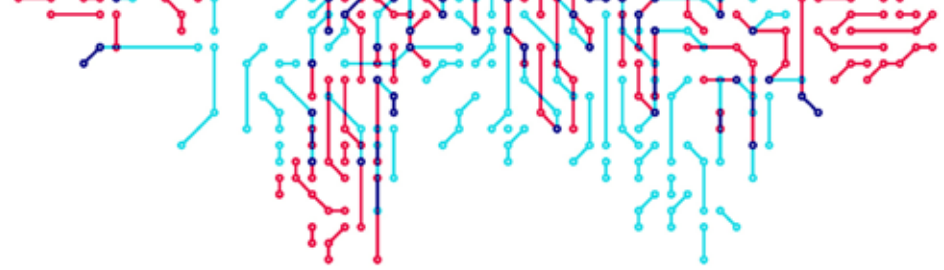
Although it is not possible to rule out that an ADA's bias is intentional in nature, it can be said that programmers, designers and responsible organisations are generally not motivated by ill will in this regard. It is more a question of ignorance or a lack of diligence. The main cause is the failure of stakeholders to have an ethical approach to computing.

An algorithm's biases may arise from the use of erroneous or biased data. As previously mentioned, the problem is that the training of machine learning algorithms requires data, and unfortunately many historical data are biased. We need to be aware of this reality and use techniques to minimise its impact.

Biases may also originate from a poorly designed algorithm. Algorithms are commonly designed with a majority group of people in mind. Although the algorithm may be accurate with respect to this group, this may not be the case for minority groups. Minority groups would thus be treated differently, hence creating a risk of discrimination. If the importance of such groups in the data set is small, discrimination may have little overall impact and go unnoticed.

A basic technique for designing unbiased algorithms is to avoid the use of potentially discriminatory variables such as gender, race, etc. This is not always possible as some





algorithms may legitimately use these variables. For example, if it has been shown that people of a particular race are more likely to suffer from a certain disease, it would be irrational not to use this knowledge. However, eliminating these variables is no guarantee against discrimination. For example, a correlation may exist between particular areas of a city and the nationality of individuals. In such a case, where someone lives would be a potential source of discrimination.

More sophisticated techniques are available to detect and mitigate the risk of discrimination. The amount of research in the field of anti-discrimination has increased in recent years. However, the effective use of these techniques comes with its own problems. For example, it is necessary to determine which groups of people are at risk of discrimination; otherwise, there is no way to know whether discrimination is occurring or how to mitigate it. The fact that these groups are often defined on the basis of data from special categories (which are subject to additional processing constraints) makes it difficult to implement these techniques.

The principle of purpose limitation

The principle of purpose limitation specifies that data should be collected for a specific and explicit purpose and should not be used in a way that is inconsistent with it. This principle is essential for people to be able to effectively control the use of their data.

Some artificial intelligence algorithms are trained for a specific purpose; e.g. deciding whether to give a loan, disease detection, etc. In such cases, the learning goal is clear. However, other algorithms fall into an “unsupervised learning” category and extract specific information from data (patterns, correlations, etc.). As the nature of this information is not known in advance, it is difficult to ensure compliance with the principle of purpose limitation.

Despite the ease with which data may be collected nowadays, AI’s great hunger for data leads its promoters to welcome the use of existing datasets. This measure may reduce costs and even enable access to certain types of data that may be difficult to collect. The reuse of data collected by a third party for a different purpose may conflict with the GDPR’s principle of purpose limitation.

The sectors that advocate flexibility in regards to the reuse of data argue that a reduction in available data may lead to less accurate and more biased artificial intelligence.

The principle of purpose limitation has some exceptions. In particular, Article 5 of the GDPR states that processing data for scientific or historical research or for statistical purposes is compatible with the initial purpose of data collection. Therefore, scientific research is a possible purpose for which AI data could be reused. The question would

thus be: what is scientific research? Recital 159 of the GDPR states that processing for scientific research purposes should be interpreted in a broad manner which includes technological development, new technology demonstration, fundamental research, applied research and privately funded research.

The principle of minimisation

The principle of minimisation states that data used in a processing operation must be adequate, relevant and limited to what is strictly necessary to achieve the purpose of processing.

As in the case of the principle of purpose limitation, the difficulties that this principle may bring with respect to ADAs are associated with the fact that these algorithms need a large amount of data to learn and make intelligent decisions. If we want a child to recognise a car, we explain its main characteristics and show them some pictures. To make a machine learning algorithm recognise cars accurately, we need to train it using many examples.

According to the principle of minimisation, only data that is strictly necessary should be processed. At the same time, the effect that a reduction in data may have on the system's accuracy and on the emergence of biases must also be taken into account.

There are several techniques that may help reduce the use of personal data:

· Feature selection. The accuracy of some machine learning techniques is highly dependent on the features involved. In such cases, adding irrelevant features may be counterproductive. Other techniques such as deep learning are not as sensitive to the features under consideration. In general terms, the selection of relevant features results in simpler and easier to train models.

· Federated learning. Suppose we want to train an artificial intelligence model on data from a group of people. Instead of giving the data to an organisation that will do the training, federated learning entails doing the training in a distributed manner: each person does the training themselves using their own data. This is achieved by giving the current model to each person, who will update it with their data and then return what they have updated. After the updates from all users have been collected, they will be combined to generate the trained model.

· Anonymisation and use of synthetic data. The aim is to avoid the use of personal data in training, irrespective of whether the data are anonymised (when the link with the person who originated the data has been broken) or synthetic (when invented data mirrors the characteristics of original data).

While an abundance of data benefits ADAs, it is equally important for the data to be accurate. Therefore, having smaller quantities of data may be preferable if they are more accurate. This notion is consistent with the principle of accuracy.

As previously stated, it is not always possible to clearly define the purpose. Adherence to the principle of minimisation is determined in terms of the end purpose. Hence if we are unable to clearly define the purpose, we cannot determine whether the principle of minimisation is being followed.

Analysis of ADA uses

To conclude this section, we will review some ADA uses presented in the fieldwork study from the perspective of data protection. For this analysis, uses that have a significant impact on people have been chosen.

Assessment of the risk of violence

The chapter on the uses of artificial intelligence in the justice system mentions several tools designed to estimate the risk of reoffending.

· RisCanvi: estimates the risk of a person reoffending after being released from prison.

· SAVRY: estimates the risk of youth reoffending.

· VioGen: estimates the risk of reoffending in cases of violence against women.

The assessment of the risk of reoffending is a necessary part of deciding which preventive or protective measures should be taken. Before the introduction of these systems, the risk of reoffending was assessed in an unstructured way by professionals with specialised knowledge such as psychologists, psychiatrists and social workers. The problem with such evaluations is that they are highly dependent on the person making them and hence on factors of importance to them, their biases, their mood, etc. Studies show that these risk estimates have very low accuracy rates.

Tools such as RisCanvi, SAVRY and VioGen are designed to reduce the dependency on the person doing the assessment. Based on an analysis of reoffending data, these tools identify the characteristics that increase risk and the interrelationship between them. After they have been identified, they are used in all risk assessments. The fact that these characteristics are public gives the system a good level of transparency. Transparency encourages the non-use of characteristics that may be an obvious source of discrimination against minorities (e.g. skin colour or country of origin).

Estimates provided by these systems are not definitive. Staff may change them if they feel compelling reasons exist to do so (specific situations of importance to an individual that are not considered within the system parameters). However, any change in assessments made by the system must be justified.

Because decisions taken by the system in such cases may significantly impact individuals, the GDPR asserts the right of individuals to demand human intervention, express their point of view and challenge decisions.

Scoring systems

Before giving a loan, a bank studies the level of risk involved. The same thing happens when you apply for insurance, such as car insurance. Nowadays, it is not bank or insurance company staff who determine the level of risk but rather an algorithm.

Banks have highly detailed profiles of their customers: they know where they work, their salary, what they spend their money on, etc. Scoring systems can use all this information to determine the level of risk associated with a loan. For example, card transactions have a code that identifies the type of activity for which the cards are used: charges from pharmacies, gambling establishments, dating services, etc. Some of these categories may increase a person's risk profile. Some algorithms make use of demographic data, social media data, etc.

By accepting the data protection clauses, customers consent to a wide range of uses including the use of their data for marketing purposes (offering products and services). Unfortunately, very few people read these clauses.

Even if they have a legal basis, the fact that the algorithm is a black box makes it difficult to determine whether such systems are fair. Some studies suggest that apparently insignificant changes (such as changing the resolution of a mobile phone) may affect outcomes. Other studies suggest the systems may perpetuate historical discrimination relating to credit access.

Given the importance of such decisions, the GDPR stipulates that individuals have the right to demand human intervention, express their point of view and challenge decisions.

Credit card fraud detection

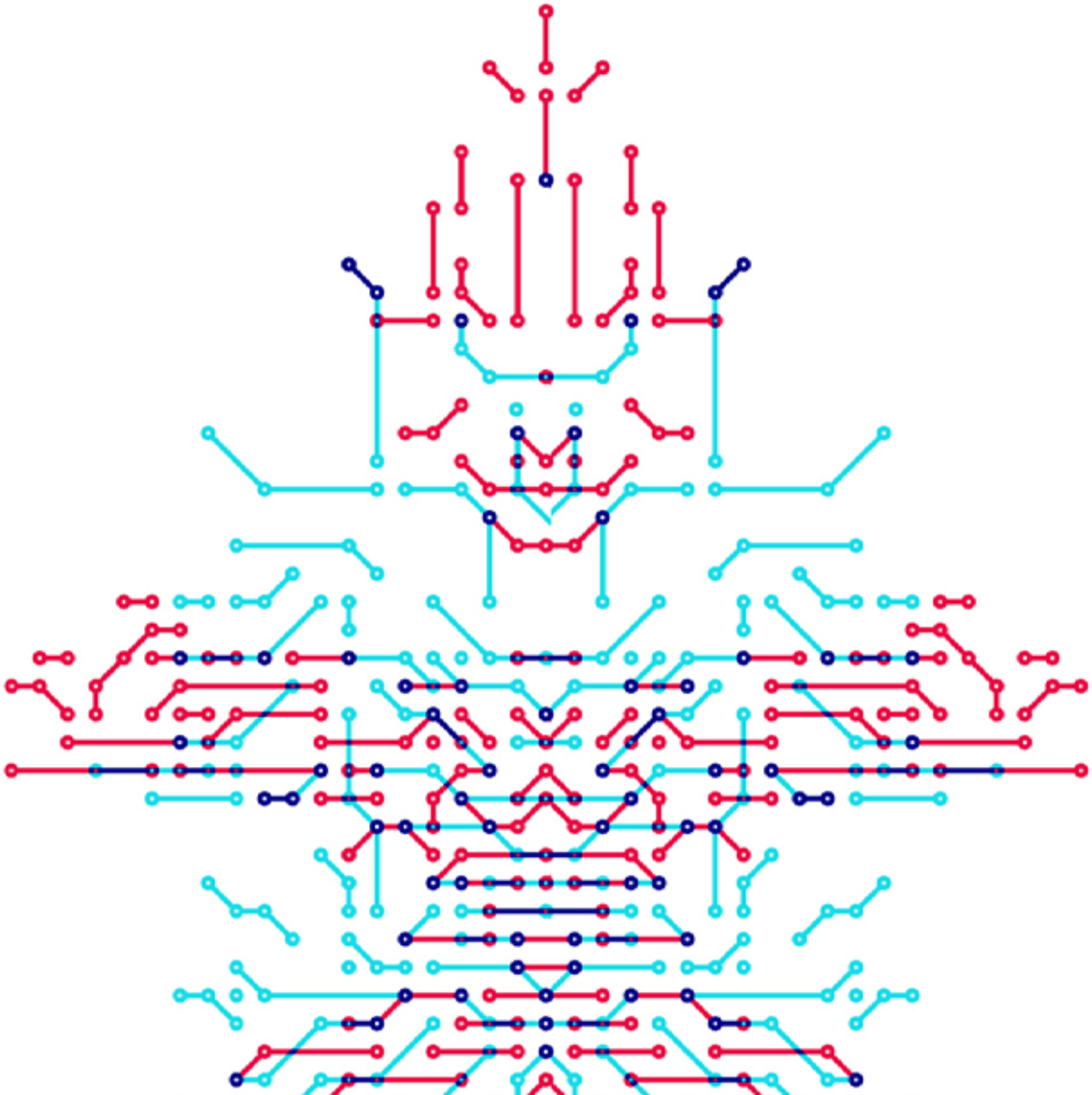
Fraud detection systems are undoubtedly designed to benefit both banks and their customers. They analyse the characteristics of purchases to ascertain whether they are fraudulent. The problem with these systems is that they have a very high false positive rate.

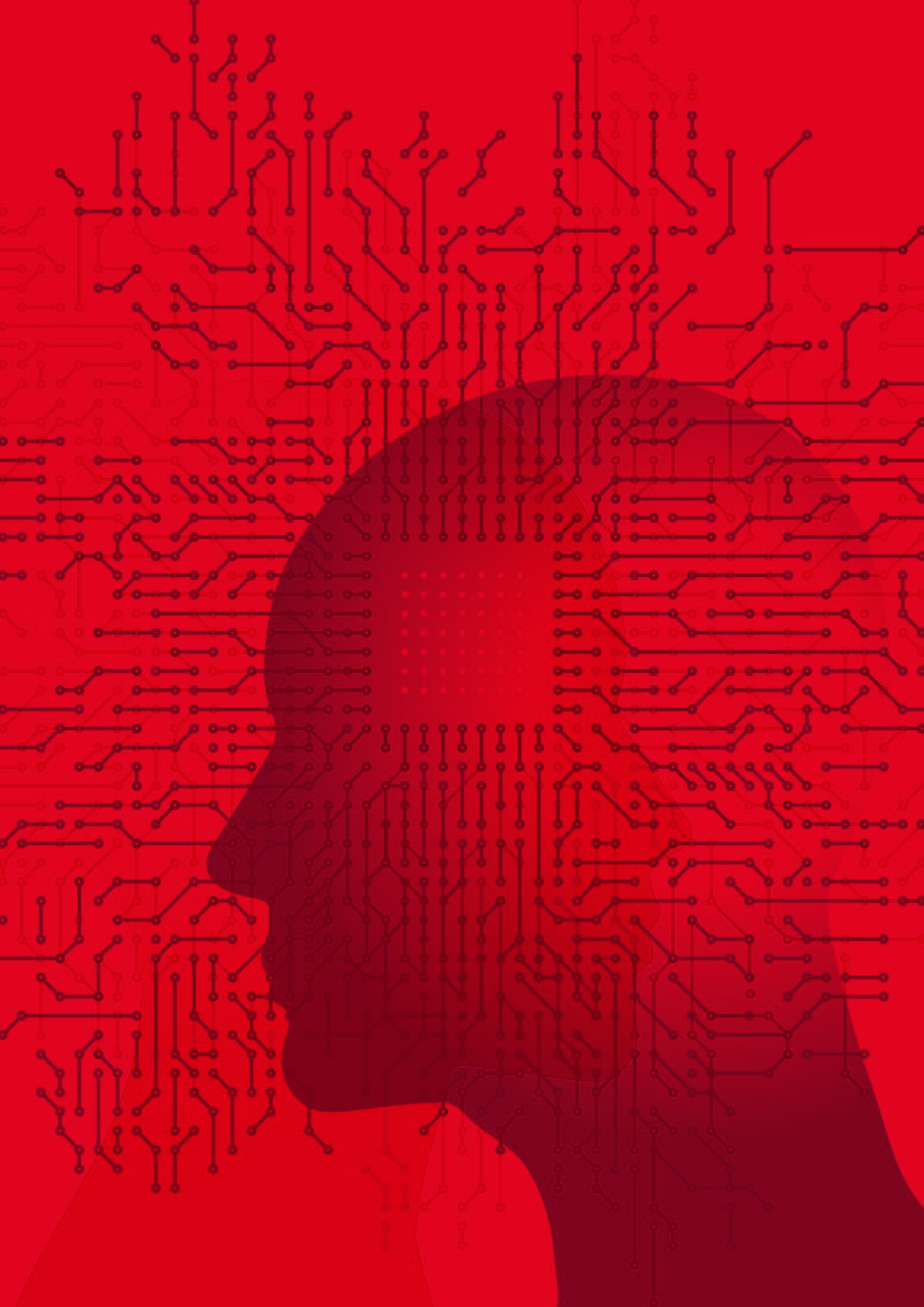
Even though the consequences of false positives are not usually serious, mechanisms should be put in place to challenge decisions.

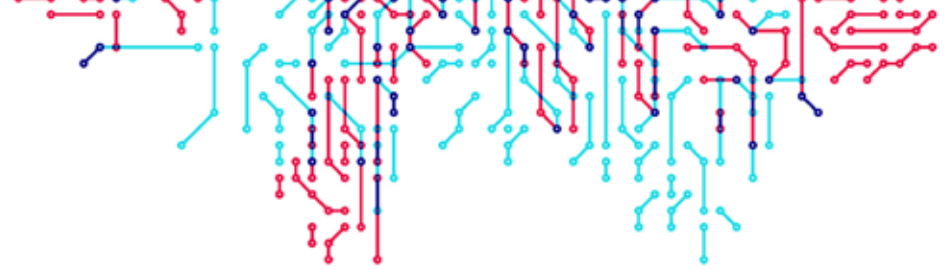
Disease detection

In medicine, disease detection is one of artificial intelligence's major uses. Three cases have been presented in the chapter on healthcare uses: the detection of colon cancer, exaggerated pain and cirrhosis.

Depending on the specific use, artificial intelligence may provide several benefits: improved detection accuracy, earlier detection, the use of less invasive techniques, etc. However, the impact that these uses may have on people's lives makes the participation of doctors, who have the last word, essential. Fortunately, in the field of medicine these systems are being proposed primarily as decision-making support systems.







4.
FINAL RECOMMENDATIONS

The ability to make decisions autonomously has been viewed circumspectly since the beginnings of AI. In 1942, an ethical code for the field was suggested in the short story *Runaround* by science fiction author Isaac Asimov. This was fourteen years before John McCarthy coined the term *artificial intelligence* in 1956.

The widespread growth that ADAs and AI have achieved in recent years, along with their potential impact on people's lives, has led businesses and other organisations to develop numerous codes of ethics. For example, the OECD has proposed a code of ethics based on the following principles:

- AI should benefit people and the planet by contributing to inclusive growth, sustainable development and wellbeing.
- AI systems should be designed to respect the law, human rights, democratic values and diversity. They should also include the mechanisms needed to ensure a just society.
- Transparency must exist to ensure that people understand the results of AI systems and have the opportunity to oppose them.
- AI systems must operate robustly and safely throughout their entire lifecycle, and potential risks must be assessed and managed on an ongoing basis.
- Organisations and individuals that develop, deploy or operate AI systems are responsible for their proper functioning, in accordance with the previously mentioned principles.

Our aim in this section is to expand on the personal data protection section that has been included in many of the abovementioned ethical codes. We accomplish this by putting forward a series of recommendations for implementation by stakeholders with agency in ADA/AI systems (individuals, organisations and supervisory authorities).

Recommendations for individuals

The GDPR is advanced data protection legislation whose purpose is to give people control over the use of their data. Yet in order for the GDPR to be effective, it will be nec-

essary for people to change their mindset. This will necessarily entail creating a culture of privacy. Fortunately, the ignorance about this issue that existed ten years ago is no longer present. While before we were delighted to get free services on the internet, we now know they were not free because we paid for them with our data. Unfortunately, we are still too permissive when it comes to the use of our data. Most people accept terms of use without having read them.

In this regard, our recommendations to people concerning the use of their data for artificial intelligence purposes are rather general in nature:

- Be aware that the personal data which an organisation processes do not belong to the organisation; it is information that has been entrusted to them by individuals.
- We need to know what our rights are when it comes to our data.
- In circumstances where our rights over our data are not respected, we must know what mechanisms are available to enforce them.
- We must develop a critical mindset when it comes to giving permission for our data to be processed. In particular, we should understand the purpose of the processing, know what data are needed, understand the potential consequences of the processing and make consistent decisions in regards to consent. For example, a gym customer may conclude that having to provide a fingerprint to enter the facility is unreasonable.
- Be aware that ADA/AI algorithms may be particularly intrusive. Once we know that our data may be processed by an ADA/AI algorithm, it is advisable to analyse the potential consequences.

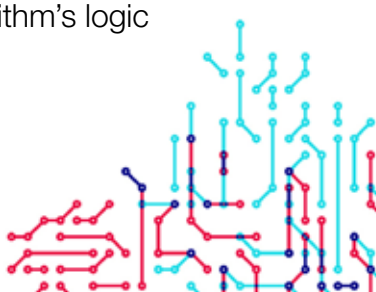
Recommendations for organisations using AI

Artificial intelligence and ADAs offer a competitive advantage that organisations should not overlook. However, their misuse may bring major consequences. For example, Cambridge Analytica was a political consulting firm that unlawfully used personal data to create personalised election advertising that emphasised aspects of political programmes which most matched voters' preferences. As a result of this scandal, Cambridge Analytica ceased to exist and Facebook (the source of the data) has had to pay a $5 billion fine.

Automated decision-making algorithms must be used **ethically** and **according to the GDPR's principles** and the individual rights that it recognises. Although the design and development of automated decision-making algorithms is primarily a technical task,

those responsible for them should be required to have basic knowledge of ethics and data protection. Existing codes of ethics need to be evaluated and new ones developed.

- The impact that AI has on people must be assessed. Automated decision-making algorithms are an innovative technology. The use of this technology is a factor that must be taken into account when deciding **whether a Data Protection Impact Assessment** (DPIA) is needed. Therefore, processing of this type is more likely to require a DPIA. In addition, implementation of the DPIA by the controller may help to demonstrate that personal data have been processed responsibly in the event that a complaint is brought by the data protection authority.
- **The right not to be subject to automated decisions**. According to Article 22 of the GDPR, in cases where the decision has significant consequences for individuals, no ADA should be implemented purely on an automated basis.
- Where the possibility of automated decision-making is not ruled out, the organisation should consider other less invasive ways of achieving the intended purpose.
- The **principle of fairness** stipulates that data use should be within reasonably expected parameters and that its consequences for individuals should not be unjustified. In the case of automated decision-making algorithms, algorithm bias is particularly important and may lead to discriminatory decisions. It is very difficult to ensure that an algorithm is bias-free, but some techniques may help to minimise bias:
  - Use unbiased data. The use of biased data in the training of an automated decision-making algorithm is one of the main ways that bias enters algorithms' decision-making processes. Training data must be analysed to detect and mitigate bias.
  - Avoid using characteristics that may lead to discriminatory decisions: sex, age, skin colour, etc. It should be borne in mind that this is not always possible and that not using these characteristics is no guarantee that indirect discrimination will not occur based on other characteristics that are a consequence of the former ones.
  - Analyse an algorithm's results with a view to detecting its possible discriminatory effects.
- The **principle of transparency** requires that individuals receive clear information regarding the processing of data. In the case of automated decision-making algorithms that have a significant impact on people, information on the algorithm's logic must be provided. With a view to fostering an algorithm's transparency:



- It is advisable to use algorithms that may be reviewed externally.
- Whenever possible, it is advisable to use explainable algorithms rather than algorithms that work like a black box.
- The information provided on the algorithm's logic must be clear and include the basic aspects of its operation.
- The **principle of data minimisation** results in a limitation of the data available within a field that makes extensive use of data. To minimise the impact of these principles, it is advisable to use techniques that reduce the amount of data that are needed. This type of data includes those discussed in the previous section: anonymisation, synthetic data generation, feature selection and federated learning.
- The principle of purpose limitation also results in a limitation of available data. In spite of this limitation, it must be ensured that the use of data is not incompatible with the purpose for which they were collected.
- In situations where the controller identifies a compatible purpose, data subjects must be told in order to give them an opportunity to make informed decisions about the use of their data.

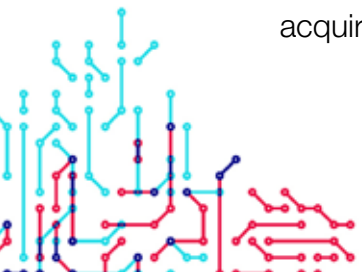
Recommendations for supervisory authorities

Conflicts between the GDPR's principles and artificial intelligence algorithms, as well as greater use of these algorithms, increase the likelihood of complaints being filed with data protection authorities charged with enforcing the regulation.

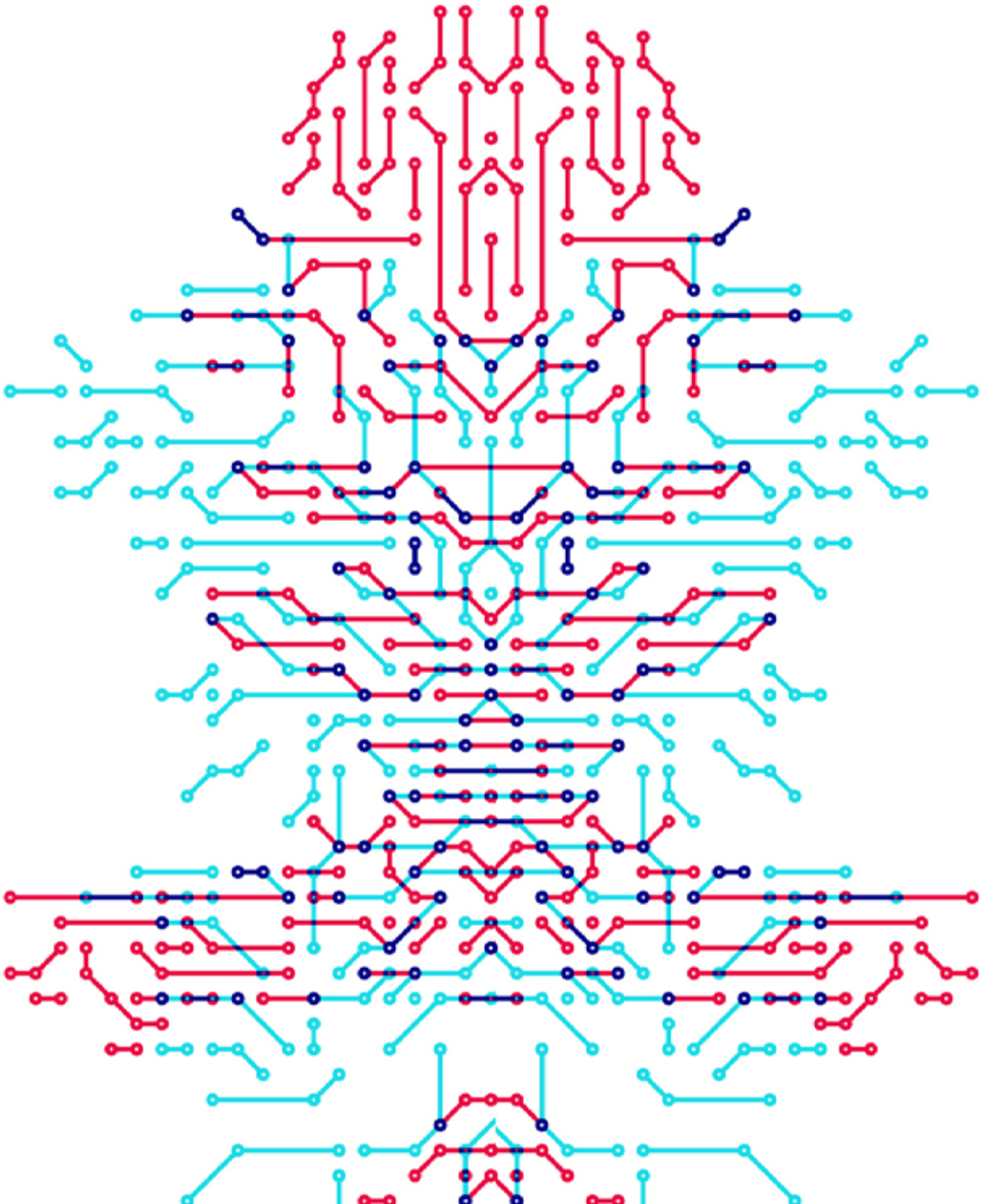
In their investigative capacity, data protection authorities may experience difficulties arising from the complexity of automated decision-making systems and artificial intelligence. For example: when assessing whether the system discriminates (fairness principle); when assessing the need to process a particular type of data (minimisation principle). The help of experts may be needed in the performance of such tasks.

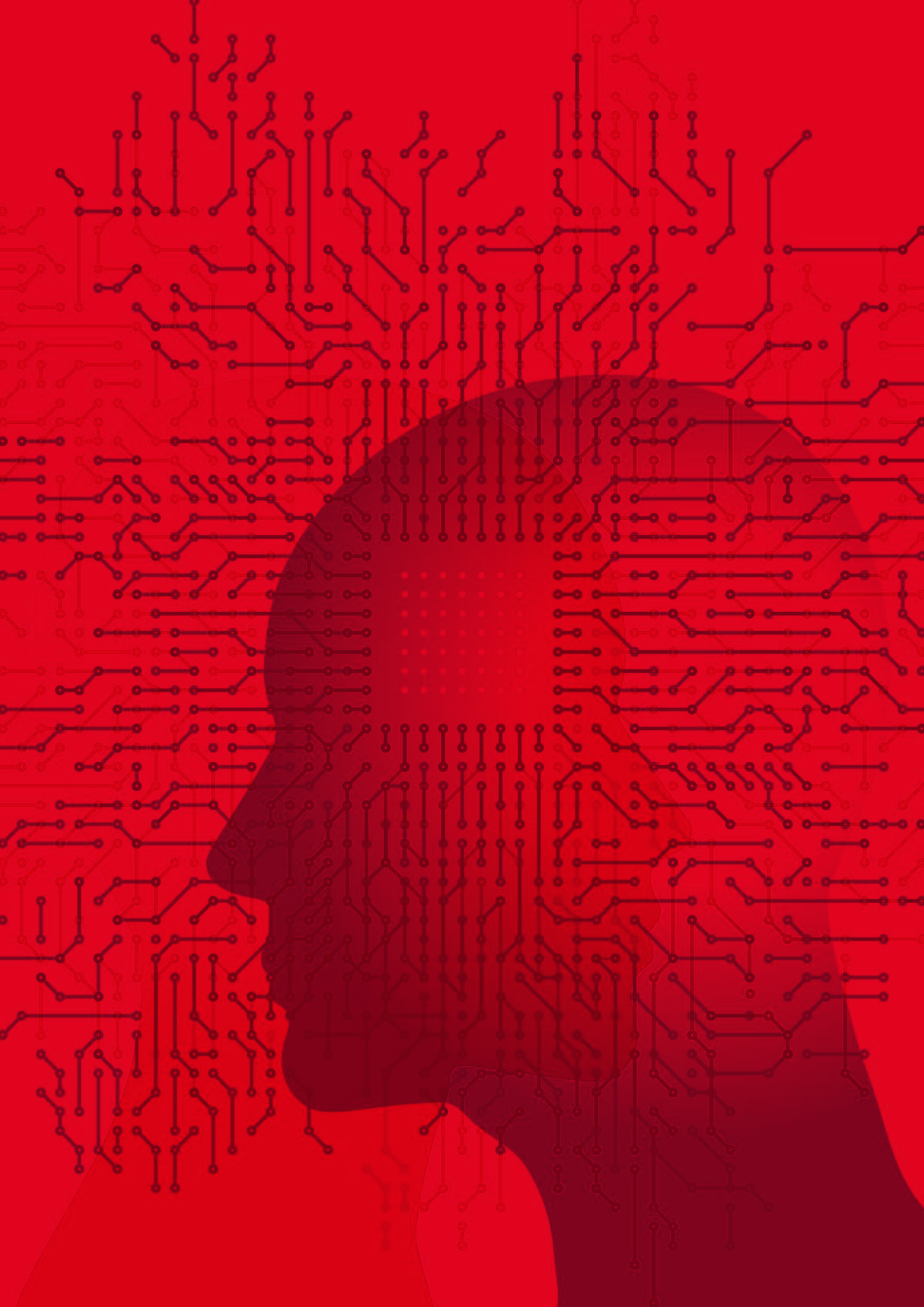
Given the importance of supervisory authorities ensuring compliance with the GDPR and requesting changes to processing actions in cases of non-compliance, the following will be essential:

- Having the knowledge needed to effectively carry out the supervisory tasks that they are assigned. ADAs, AI and technology in general are part of a rapidly evolving field. Data protection authorities must have sufficient knowledge of it or the ability to acquire it when needed.



- Efforts should be made to raise public awareness of both the rewards and the risks associated with ADAs and AI.
- Efforts should be made to raise awareness among organisations that use ADAs and AI about the problems associated with data protection and encourage the inclusion of such issues in codes of conduct.





5.
GLOSSARY

A

Algorithm. A precise sequence of instructions for the performance of a given task. Algorithms are everywhere. For example, the steps needed to enrol in a school or apply for social assistance can be considered to be algorithms. Algorithms are essential to computing. In fact, the only thing a computer does is follow the instructions of the algorithms with which it has been programmed.

Anti-discrimination. A set of techniques used to evaluate and mitigate processing that discriminates against certain groups of people.

Artificial intelligence (AI). A generic term referring to algorithms designed to give computers intelligent behaviour. The definition of *intelligent behaviour* is elusive. Two main types can be identified: general AI (which seeks to give computers human intellectual capabilities) and weak AI (which focuses on specific tasks such as voice recognition, image recognition, etc.). Artificial intelligence has surpassed human capabilities to perform many specific tasks, although general AI is still far off.

Automated decision-making algorithm (ADA). An algorithm that makes automated decisions without human participation. In recent years, the use of these algorithms has become widespread. For example, it is estimated that 70% of financial transactions are performed by algorithms. The potential effects of these algorithms on individuals raise important ethical issues, such as the accountability that may be associated with decisions (e.g. in an autonomous car accident), etc.

B

Black box. The inherent complexity of many models used in machine learning makes it difficult to explain their behaviour. In such cases it is said that the model behaves like a *black box*: it receives inputs and produces an output, but we do not know how the output is generated from the inputs. Each model has a specific level of transparency. For example, a system that learns rules based on training data is very transparent, whereas a system based on deep learning is a black box (e.g. the model's complexity makes it difficult to determine which input data characteristics are involved in determining the outcome).

C

Consent. The data subject's clear, informed and free expression of their acceptance to having their data processed.

Controller. A person or organisation that determines the purpose for and manner in which data is to be processed.

D

Data subject. Person to whom data refer.

Deep learning. This term refers to the use of complex neural networks that have multiple layers of neurons. The practical use of this technique is quite recent owing to its high computing cost.

F

Fairness. A key principle in the GDPR. Data processing is fair when it uses the data in a way that would be foreseeable by data subjects and when it does not give rise to unjustified consequences. For example, fair processing should not discriminate by gender, skin colour, etc.

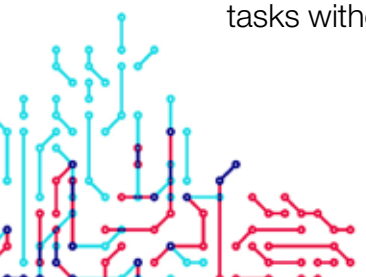
Federated learning. A technique used in machine learning to train the system in a decentralised way so that a central entity does not need access to all training data. The current model distributes the set of entities that participate in the training. Each of these entities updates the model based on its data and returns the update. Federated learning delivers two important benefits: it improves people's privacy (as no single entity has access to all data) and it distributes the computing cost of training across several entities.

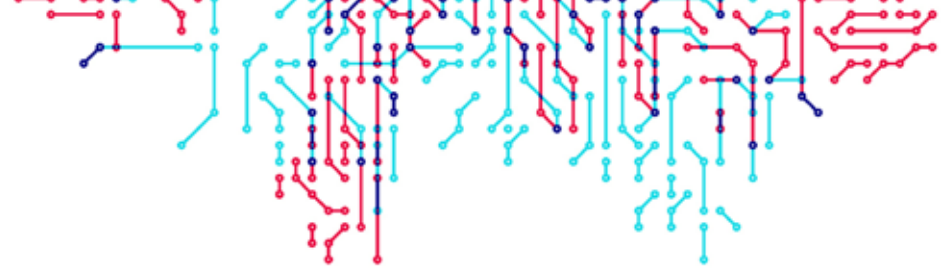
G

General Data Protection Regulation (GDPR). European regulation on the use of personal data. Having entered into force in May 2018, it replaced Data Protection Directive 95/46/EC and reinforces the safeguards it provided.

M

Machine learning. A subset of AI algorithms used to teach computers how to perform tasks without having to give precise instructions. Machine learning is particularly useful





for teaching computers to do complex tasks for which no particular algorithm is known. For example, we know of no algorithm to determine whether an email message is spam. What machine learning can do is analyse the data from emails marked as spam to determine the characteristics that are indicative of spam.

Model. A representation of an underlying reality that is used by automated decision-making algorithms to make predictions and decisions. Models are tailored to each individual case in a learning process based on available data.

N

Neural network. A machine learning model inspired by the functioning of the neuron networks in living organisms.

P

Personal data. Data that are or can be associated with a person.

S

Supervised learning. A technique used in machine learning to train the system. An algorithm that uses a mathematical model to produce a result based on input data. Based on a set of training data that has the algorithm's expected inputs and outputs, supervised learning fine-tunes the model to minimise errors on the training data.

T

Transparency. A key principle in the GDPR. Transparent processing requires information to be given to data subjects in a clear and understandable manner.

U

Unsupervised learning. A technique used in machine learning to train the system. Based on a set of training data, unsupervised learning looks for patterns in these data (cause-and-effect relationships, correlations, standard groups, etc.).