

HOW CAN HUMANS KEEP THE UPPER HAND?

The ethical matters raised by algorithms and artificial intelligence

REPORT ON THE PUBLIC DEBATE LED BY THE FRENCH DATA PROTECTION AUTHORITY (CNIL)
AS PART OF THE ETHICAL DISCUSSION ASSIGNMENT SET BY THE DIGITAL REPUBLIC BILL

DECEMBER 2017

FOREWORD



The hopes and fears sparked by artificial intelligence today have raised it to mythological status in many imaginaries and narratives. Some foresee the massive destruction of our jobs, the apocalyptic emergence of a hostile robotic consciousness or the downfall of Europe, called to get overwhelmed by the global competition. Others, meanwhile, are being pinned on the dream of a tailor-made world, a new Golden Age where every thankless or repetitive task would be delegated to machines; an Eden where infallible tools would have put an end to illness and crime, and perhaps even political conflict; in short, evil would have become a thing of the past. Under its myriad guises that are, by turn, fascinating or alarming, bright or gloomy, AI arguably says more about our fantasies and fears than about what our future world will look like. The appeal of this type of eschatological rhetoric in Europe is such that it shows to what extent technology allows visions of the future and generates a power to look ahead, which is sometimes lacking from our political visions.

Deflating all this hype surrounding new technology is one thing. It does not mean that we are unaware of the many upheavals and new challenges to be addressed as these unprecedented tools or assistants invade every corner of our lives. Safeguarding our autonomy in human decision-making processes in the face of machines that are sometimes considered infallible, detecting discrimination generated unintentionally by scalable systems, protecting collective processes that are in some instances being undermined by the sheer power of digital personalisation... the challenges are multiple and their implications already tangible, prompting questions over some of the major social pacts on which our community life is founded.

The priority for public authorities should be to clearly identify what these challenges are. It is only then that suitable solutions may be found, to build technological innovation into a well-defined vision of our future. This was the idea behind the assignment to reflect on the ethical issues raised by digital technologies, which the Digital Republic Bill set to the French Data Protection Authority, the CNIL in 2016.

How should such an assignment be understood and undertaken? Many people have pondered over this, and even questioned this new responsibility of the CNIL. How can ethics be grasped and stated with regard to highly complex and changeable subjects, in what capacity, according to what approach?

By seeking to think on the fundamental principles underpinning the life of humans and societies, and thus shaping a shared social pact on a complex matter at a given time, ethics is an eminently collective, plural concept. In the very specific field of health and life sciences, the composition as well as the collegiality principle governing the work of the French governmental advisory council on bioethics issues (CCNE) meet this need for plurality.

As custodian of the ethical principles set by the lawmaker forty years ago, the CNIL is in a legitimate position to lead this ethical discussion, at a time when new technical possibilities are raising new challenges or calling the status quo into question.

However, it was clear that the CNIL could not lay claim to any sort of monopoly over ethical discussions on digital technology. On such a vast and cross-cutting subject, on no account should these be held behind closed doors. Digital technology is not a sector, which could be entrusted to a select ethics committee of just a few members – however competent they may be. So we had to innovate.

And it was in this mindset that the CNIL set a collective approach in motion, for several months overseeing a public debate with the help of partners from various sectorial fields (health, justice...). In this respect, ethics is just as much about the process itself as it is about the outcome. We thus decided to take as our starting point the uses, existing questions and possible solutions outlined by the participants in the debate. This report and the recommendations it contains have been written on the basis of the comments and viewpoints voiced at more than forty events held in Paris and across the rest of France.

Innovative public action was also required to accommodate the need to involve the general public more closely in the public discussion since this complex world is modelling its existence to an ever greater extent and entails fundamental societal choices. A world in which we are all called, citizens included, to play an increasing part. The CNIL therefore organised a public consultation day in Montpellier, on 14 October 2017, at which some forty volunteers were able to add their voices to the polyphony of the public debate.

The first benefit of this inclusive, decentralised approach is to have opened the debate up as widely as possible, thus helping to enhance French society's knowledge on the questions raised by algorithms and AI. This appears as crucial since limiting the debate to a few expert circles would risk arousing mistrust and suspicion, particularly regarding the increasing complexity and fragmentation of socio-technical systems, and the not always clearly foreseeable impacts of constantly evolving artefacts. Enabling all of our fellow citizens to become informed and critical users of technologies is of the utmost importance ethically, democratically and pragmatically speaking. For the CNIL, this is also a way of continuing to assist in familiarising French society with digital technology, a mission it has been accomplishing for 40 years.

At a time when France is setting out its position regarding artificial intelligence, with Europe due to do the same thing shortly, the report following these several months of public debate is helping to lay the groundwork for a collective thought process. It provides an overview of the issues and outlines a certain number of principles and recommendations.

What these all have in common is to enable humans to “keep the upper hand”. Amid broad-scale digitisation, this may seem somewhat out of step with reality. And yet we believe that this is precisely where we should be firmly focusing our joint attention. On making sure that these new tools are designed by humans, to serve humans, in a mindset of transparency and accountability.

May these discussions fuel those under way within the public authorities, including the Villani mission's, as well as within the various sections of civil society. In this way, may they help to shape a French model for the ethical governance of artificial intelligence.

CONTENTS

EXECUTIVE SUMMARY	5
AN INNOVATIVE APPROACH TO CRAFTING A COLLECTIVE AND PLURALIST ETHICAL THOUGHT PROCESS	7
KEY DATES AT A GLANCE	10
KEY FIGURES AT A GLANCE	11
ALGORITHMS AND ARTIFICIAL INTELLIGENCE TODAY	13
More precise definition is required to clarify the public debate	14
Algorithms: a central feature of computer science with a long history	15
From algorithms to artificial intelligence	16
Steer the discussions towards the most crucial impacts and applications of algorithms today	19
Use and promise across all sectors	21
THE ETHICAL ISSUES	23
Ethics, as a prefiguration of legal standards	24
Autonomous machines: a threat to free will and responsibility	26
Bias, discrimination and exclusion	31
Algorithmic profiling: personalisation versus collective benefits	34
Preventing massive files while enhancing AI: seeking a new balance	38
Quality, quantity, relevance: the challenges of data curated for AI	39
Human identity before the challenge of artificial intelligence	41
HOW CAN WE RESPOND?	43
From ethical thinking to algorithmic regulation	44
What the law already says about algorithms and artificial intelligence	45
The limits of the current legal framework	46
Should algorithms and artificial intelligence be banned in certain sectors?	47
Two founding principles for the development of algorithms and artificial intelligence: fairness and continued attention and vigilance	48
Engineering principles: intelligibility, accountability, human intervention	51
From principles to policy recommendations	53
CONCLUSION	61
ACKNOWLEDGEMENTS	62
LIST OF EVENTS ORGANISED FOR THE PUBLIC DEBATE	63
GLOSSARY	66

EXECUTIVE SUMMARY

This report is the result of a **public debate** organised by the French Data Protection Authority (CNIL). Between January and October 2017, **60 partners** (associations, businesses, government departments, trade unions, etc.) held **45 events across France** with a view to identifying the ethical concerns raised by algorithms and artificial intelligence, as well as possible solutions for addressing them.

Part One of the report provides a **pragmatic definition of algorithms and artificial intelligence (AI)** and presents their main uses – with particular emphasis on those that are most in the public limelight today. Traditionally, an algorithm is defined as a finite and unambiguous sequence of instructions for producing results (output) from initial data (input). This definition covers the multiple digital applications which, by using programs that themselves translate algorithms into a computer language, fulfil such diverse functions as yielding results on a web search engine, providing a medical diagnosis, driving a car from A to B, detecting fraud suspects among social welfare recipients, etc. In contemporary public debate, artificial intelligence mainly refers to a new class of algorithms, configured on the basis of machine learning techniques. The instructions to be carried out are no longer explicitly programmed by a human developer; instead they are generated by the machine itself, which “learns” from the input data. These machine learning algorithms can perform tasks that traditional algorithms were incapable of doing (picking out a particular object from vast image datasets for example). But their underlying logic remains incomprehensible and a mystery even to those who wrote them.

The public debate has highlighted 6 main ethical issues:

- The development and growing autonomy of technical artefacts are paving the way for ever more complex and critical decisions and tasks to be delegated to machines. In these conditions, whilst technology may well be increasing our capacity to act, is it not also posing a **threat to our autonomy and free will**? Will the prestige and trust placed in machines, often assumed to be “neutral” and fail-proof, tempt us to hand over to machines the burden of responsibility, judgment and decision-making? How should we tackle the ways in which complex and highly segmented algorithmic systems might end up watering down responsibilities?
- Algorithms and artificial intelligence can create **bias, discrimination and even exclusion**. Although such

phenomena can be intentional, a far more pressing matter at a time when the development of machine learning algorithms is upon us, is their development without us even knowing it. How should this challenge be addressed?

- The digital ecosystem as it has grown with the Internet, and actuarial techniques before that, have largely tapped into the possibilities offered up by algorithms in terms of personalisation. Individuals have gained a great deal from profiling and ever finer segmentation, but this mindset of personalisation is also likely to affect not just individuals but also the key **collective principles forming the bedrock of our societies** (democratic and cultural pluralism, risk-sharing in the realm of insurance).
- By being grounded in machine learning techniques, artificial intelligence requires vast amounts of data. And yet we are all too aware of the risks the creation of such massive files poses for our personal and public freedoms. The data protection legislation therefore advocates an approach where the **collection and retention of personal data** should be minimised. Does the promise held by AI justify to rethink the balance?
- The **choice of which and how much data should be used by an algorithmic model**, and thus the existence of potential bias in datasets curated to train algorithms, are of paramount importance. This matter calls us all for a critical attitude so as to avoid placing excessive trust in machines.
- AI involves an increasing autonomy of machines and the emergence of forms of **hybridisation between humans and machines** (hybridisation both in terms of action assisted through recommendation, and very likely in physical terms in the future). This challenges the notion of our human uniqueness. Should we and, indeed, is it possible even, to speak in literal terms of “ethics of algorithms”? How should we view the new class of objects, humanoid robots, which are likely to arouse emotional responses and attachment in humans?

Part Three of the report considers the possible responses outlined during the public debate.

It looks firstly at the principles that are likely to frame an AI that benefits and empowers humans. **Two new founding principles have come to light.**

The first, substantial one, is the **principle of fairness**. It builds on the principle initially proposed by the French

Council of State on digital platforms. A fair AI also considers collective outcomes, meaning that the algorithmic tool cannot betray its community of users (who might be consumers or citizens), whether or not it processes personal data.

The second, more methodical, principle is that of **continued attention and vigilance**. It seeks to address, over time, the challenge posed by the unstable and unpredictable nature of machine learning algorithms, as well as to provide an answer to the forms of indifference, negligence and watering down of responsibility to which highly segmented AI systems can give rise. Lastly, it is aimed at taking on board and offsetting the form of cognitive bias that leads to us placing excessive trust in the prescriptive statements of algorithms. The point is to organise, through specific measures and procedures, an ongoing, methodical, deliberative and productive thinking process on these machines. This should involve all the stakeholders throughout the "algorithmic chain", from the designer to those who train algorithms to the end users.

Both of these principles appear to underpin the regulation of the complex assistants and tools that AI and algorithms represent. They not only allow for their use and development, but also their oversight by the community.

They are rounded off by a discussion **on two other engineering principles** which are particularly relevant when

it comes to AI: one aimed at rethinking the requirement for human intervention in algorithmic decision-making (Article 10 of the French Data Protection Act); the other at organising the intelligibility and accountability of algorithmic systems.

These principles are then set out in the form of **6 practical policy recommendations** intended for the public authorities as well as the general public, businesses and associations, for example:

- Fostering education of all players involved in the "algorithmic chain" (designers, professionals, citizens) in the subject of ethics;
- Making algorithmic systems understandable by strengthening existing rights and organising mediation with users;
- Improving the design of algorithmic systems in the interests of human freedom;
- Setting up a national platform for auditing algorithms;
- Increasing incentives for research on ethical AI and launching a participatory national worthy cause on a general interest research project;
- Strengthening ethics within businesses.

An innovative approach to crafting a collective and pluralist ethical thought process

A national public debate on the ethical matters raised by algorithms and artificial intelligence

The 2016 Digital Republic Bill gave the French Data Protection Authority (CNIL) the assignment of leading discussions on the ethical and societal matters raised by the rapid development of digital technologies.

In 2017, the CNIL decided to focus these discussions on algorithms in the age of artificial intelligence. Unbeknownst to us, these are increasingly creeping into every corner of our lives: web search engine results, financial orders placed by robots on the markets, automated medical diagnoses, allocation of students applying to universities. Across all these areas, algorithms are at work. In 2016, the subject of algorithms rushed in an unprecedented manner into the public debate scene, garnering widespread media coverage (questions over the algorithm of the university admissions online portal "Admission Post-Bac", the use of artificial intelligence in Trump's election campaign strategy, the role of social media in the spread of fake news).

Ethical thinking concerns decisive societal choices. It should not proceed without taking this pluralist and collective dimension into account – especially when it deals with such a cross-cutting issue, with a bearing on all aspects of our social and personal lives. It would simply not be feasible to bring together within a single committee all of the expertise and perspectives necessary for examining the matters raised by algorithms in sectors as varied as healthcare, education, marketing, culture, defence and security for example.

So rather than holding centralised discussions on these subjects directly, the CNIL decided to adopt an original stance as a leader of an open and decentralised national public debate. At a launch organised on 23 January 2017, it thus called on all the interested stakeholders and organisations – public institutions, civil society, businesses – to host a debate or event on the subject, on which they would then report back to the CNIL. The aim was therefore to gather from the stakeholders on the ground the ethical subjects identified to date as well as their ideas for addressing these.

Sixty partners came forward in response to the appeal launched by the CNIL, harking from very different sectors and representing a variety of setups. Among them, we could mention the "Ligue de l'Enseignement" (association that focused on education concerns), French Insurance Federation (FFA), French Ministry of Culture (DG-MIC), Open Law (association that reflects on the justice system) as well as trade unions such as CFE-CFC and FO Cadres (for recruitment and HR), etc.

Ethical thinking concerns decisive societal choices.

It should not proceed without taking this pluralist and collective dimension into account



They organised 45 events between March and October 2017 across several French cities (as well as abroad through the Future Society at Harvard Kennedy School), in which some 3,000 people took part. The CNIL provided overall, coherent coordination of the events.

The events organised for the public debate were also an opportunity **to get French society at large thinking about issues on which awareness among all our contemporaries, not just among experts, is of the utmost democratic and civic importance.**

Public consultation: Montpellier, 14 October 2017

The questions raised by algorithms and artificial intelligence have to do with societal choices, and have implications for all citizens. A consultation was therefore organised so as to find out what the general public think about all this. The aim was to round off the views gathered during the various events, mainly from experts in the different sectors.

A consultation day was therefore scheduled on 14 October 2017, with support from the City of Montpellier and Montpellier Méditerranée Métropole. A diverse 37-strong citizen panel was formed following a call for applications.

The format adopted sought **to encourage the sharing of ideas and the formation of a collective opinion.** The procedure enabled the participants in turn to:

- Understand the basic concepts behind algorithms and artificial intelligence;
- Jointly analyse four case studies (medicine and health-care / human resources / personalisation and filter bubbles / education and transparency) to identify the threats and opportunities associated with using algorithms;
- Come up with recommendations to ensure that algorithms and AI are deployed within an ethical framework, and assess the level of consensus reached on these.

The outcomes and insights gained are presented in the insets headed "What the public think".



Structure of the report

Overviews reporting back on the events organised by the partners and the public consultation were made to the CNIL. The views of the diverse stakeholders (trade unions, associations, businesses, researchers, citizens, etc.) across a wide range of sectors (from insurance to education, justice and healthcare) thus informed the writing of this report, which provides an overview of the ethical matters raised by algorithms and artificial intelligence in their current applications and their potential uses in the relatively short term.

As well as leading the public debate, the CNIL has also been responsible for reporting back on it and, in this regard, it has had to decide on how to structure the report, which has inevitably entailed making certain choices. Precedence has been given to providing a full and fair account of all the different views expressed, and this explains why the recommendations set out at the end of the report are not intended to settle the debate so much as to leave open a certain number of options (proposals may take the form of requirements or incentives for example), where further arbitration will thus be required. The aim is therefore to inform public decision-making rather than replace it.



The CNIL also relied on documentary research to draw up this report, often initiated at the recommendation of a particular partner. The articles or publications referred to are cited in footnotes. Reference could also be made to the pages of the CNIL's website dedicated to the ethical debate to find some select bibliographic facts¹. Lastly, the findings of a certain number of studies already carried out by various institutions in France and abroad have been analyzed, including the OPECST (French Parliamentary office for scientific, and technological assessment), CERNA (Allistene's research committee on ethics), CNum (French Digital Council), French Council of State, CGE (General Economic Council), White House, "France IA", INRIA (French Institute for Research in Computer Science and Automation) and AI Now.

The questions raised by algorithms and artificial intelligence have to do with societal choices, and have implications for all citizens

¹ <https://www.cnil.fr/fr/ethique-et-numerique-les-algorithmes-en-debat-1>

KEY DATES AT A GLANCE

7

OCTOBER
2016

The “Digital Republic” Bill gives the CNIL the assignment of leading a discussion on the ethical and societal issues raised by new technology

23

JANUARY
2017

The CNIL names algorithms and artificial intelligence as its theme for 2017 and launches it with round tables bringing together experts on these subjects

END
MARCH
2017

The first events are held by the partners in the public debate

DÉBUT
OCTOBRE
2017

45 events are hosted by **60 partners** in the public debate

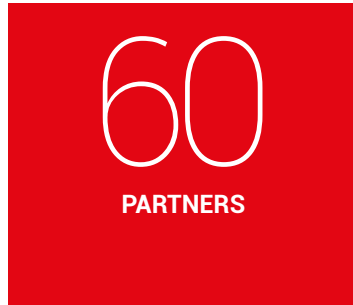
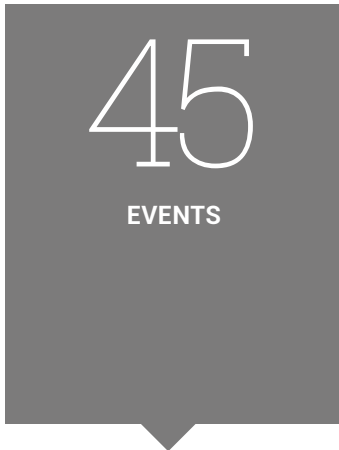
14
OCTOBER
2017

The CNIL organises a consultation in Montpellier with some **40 citizens**

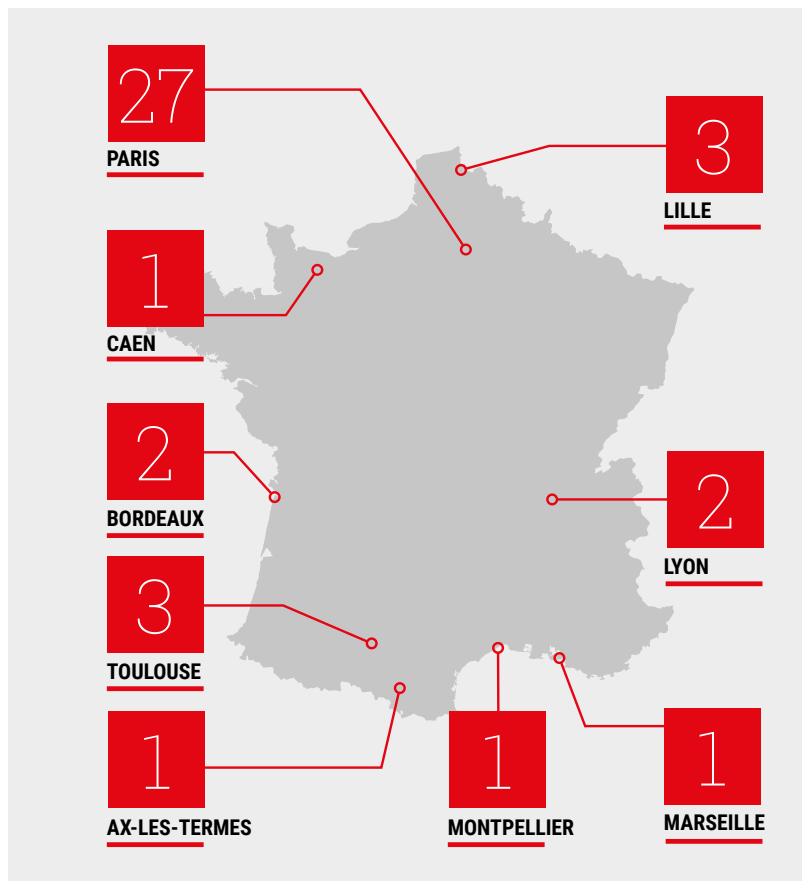
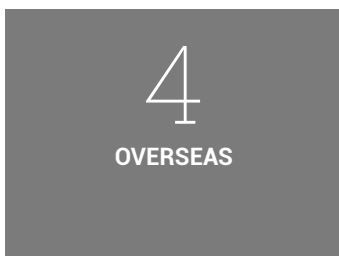
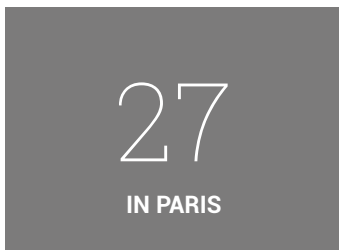
15
DECEMBER
2017

The CNIL presents the report “**How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence**”, which summarises the public debate

KEY FIGURES AT A GLANCE



NEARLY **3 000** PEOPLE ATTENDED
THE EVENTS



Algorithms and Artificial Intelligence today

**More precise definition is required
to clarify the public debate**

P.14

**Algorithms: a central feature of computer science
with a long history**

P.15

From algorithms to artificial intelligence

P.16

**Steer the discussions towards the most crucial impacts
and applications of algorithms today**

P.19

Uses and promises across all sectors

P.21

Algorithms and AI today

More precise definition is required to clarify the public debate


Algorithms and artificial intelligence are the buzzwords of the moment – but not everyone is clear on what they mean. The definitions and examples currently provided in the public debate are often fairly vague – even conflicting at times. This situation can be explained by the highly technical nature of subjects that however became topics for discussion far beyond the circles of experts and specialists to which they have traditionally been confined.

Hence, were you to actually notice, the lack of any sort of precision in the terms used. **What, indeed, do the austere notion of “artificial intelligence” defined in cybernetic circles in the 1950s and the mainstream understanding of the term, as primarily depicted by the movies in Hollywood, have in common?** And who, incidentally, pays attention to the fact that “intelligence” does not mean the same thing in French and English – the language in which the expression “*artificial intelligence*” was invented? How can we make sense of the fact that, here, we are saying that algorithms are new, while others are busy insisting that humans have been using them for thousands of years?


Beyond the realities and technical projects they are supposed to denote, algorithms and artificial intelligence have come to represent new mythologies of our time. The mere mention of which is enough to imply digital innovation and

modernity. It is hardly surprising, then, that these terms are often mistakenly associated with situations or companies eager to project an appealing and futuristic image. Presenting one’s business as coming within the realm of AI is, for many stakeholders today, a tactic to enhance their image, as it used to be a few years ago with the expression “Big Data”. Whatever, the fact of the matter is that the promise held by AI is a more or less explicit bone of contention between AI researchers, entrepreneurs and opinion leaders of all kinds within the realm of digital technologies.

Another type of confusion, which we will return to later on, seems at times to be fuelled by stakeholders whose business is generally acknowledged as coming within the sphere of artificial intelligence. The latter would appear bent on hugely overestimating not so much the opportunities as the threats of an artificial intelligence which could end up gaining such complete autonomy from its designer that the whole of humanity itself is imperilled. The most competent voices are rising up to stamp out such predictions, compared at best to fantasies and at worst to lies. These stances seem to divert public attention from the more mundane, albeit more pressing issues raised by the growing of artificial intelligence. Avoiding discrimination or protecting personal data are such down-to-earth matters.



The key to setting up a constructive discussion on the subjects of algorithms and artificial intelligence is to clearly specify the link between these two subjects



Let us be clear from the start: any definition on these subjects may appear questionable depending on the point of view. In the interests of this report, we are looking for a minimum and operational basis for discussion. This is the prerequisite to pragmatically outline the scope of the algo-

rithms and artificial intelligence systems that are raising such crucial ethical and societal issues. In other words, the point is to put forward as precise a definition as possible, but which takes into account the reasons why we care about AI today and what makes it worthy of attention.

SURVEY

Algorithms and AI: a subject poorly grasped by French citizens*

There is an awareness about algorithms in France, but a fair amount of confusion too. **83% of the population may have already heard of them, but more than half are not exactly sure what they are (52%).** 85% of French people reckon they are a massive part of everyday life, and 65% believe this trend is only set to rise further in the years to come.

83 %
of French
people have
already heard of
algorithms

* Poll carried out by the French market research and opinion poll institute, IFOP, for the CNIL in January 2017 (among a sample of 1,001 people, representative of the French population aged 18 years old and over) on the level of awareness of algorithms among the French population.

Algorithms: a central feature of computer science with a long history

In the strict sense of the term, an algorithm is the description of a finite and unambiguous sequence of steps (or instructions) for producing results (output) from initial data (input). A recipe is an algorithm for example, as a dish can be made from its ingredients². There are records of algorithms being used to solve equations dating back centuries, as early as the 3rd millennium BC in Babylon, Mesopotamia.

In our increasingly digitalised world, **computer algorithms make it possible to combine the most diverse pieces of information and to produce a wide variety of results: simulating the changing spread of flu in winter, recommending**

books to customers based on choices already made by other customers, comparing digital images of faces or fingerprints, autonomously operating vehicles or space probes, etc.

For a computer to be able to run an algorithm, it must be written in a computer language and coded into a program (a sort of text comprising written instructions, also known as “source code”). This program can then be run in a software or compiled in the form of an application. Software generally makes use of a number of algorithms: for inputting data, computing and displaying the results or communicating with other software programs, etc.

² Voir par exemple : <http://www.cnrtl.fr/definition/algorithme>

From algorithms to artificial intelligence

There are few notions whose use changes quite so much as that of “artificial intelligence” (AI) today. In this report, the decision was made to pragmatically focus on the actual uses already being made of artificial intelligence and, more precisely, on those that have developed most swiftly over recent years (in step with the progress accomplished by machine learning).

Broadly speaking, artificial intelligence can be defined as “the science of making machines do things that would require intelligence if done by men” (Marvin Minsky). The notion of artificial intelligence officially came about in 1956, during the Dartmouth Conference among cybernetic circles. Still, **Alain Turing’s** article published in 1950 (*Computing Machinery and Intelligence*) might be considered the starting point, as it was here that the latter asked the question: “Can machines think?”. Researchers in this emerging field aspired to create a general intelligence – similar to the one of humans – that could be embedded in computers. The latter would extend far beyond a limited number of fields or tasks.

Progress over the course of artificial intelligence history since the 1950s has not been continuous. To begin with, researchers have been obliged to turn their attentions away from the objective of developing artificial general intelligence (strong AI) towards more specific tasks, solving such problems as image recognition, natural language understanding or playing games (draughts, chess or Go for example). This is referred to as “weak AI”, as it is focused on one narrow task. Even if we look solely at this form of AI, the history of this research area and its applications has not all been plain sailing. A period of optimism in the 1980s gave way to an “AI winter” in the 1990s when progress faltered on account of inadequate computing power and available data in particular.

The last few years have, on the contrary, seen a series of milestones which have shone the spotlight back on the promise AI holds. Alpha Go’s (Google) victory over the Go world champion, Lee Sedol, in March 2016, was the most remarkable, in symbolic terms, of these achievements. Unlike chess, Go does not lend itself to the memorisation of a large number of moves that the machine could simply reproduce. It instead gives rise to a sheer number of possible combinations.

Alpha Go’s victory illustrates the fact that recent breakthroughs made in AI can particularly be put down to **development of the machine learning technique**, which is one of its applications. Whereas, in the past, programmers had to break down into multiple instructions the task that was being automated so that all of the steps involved were clearly specified, machine learning involves presenting the machine with example inputs of the task that we wish it to accomplish. In this way, humans *train* the system by providing it with data from which it will be able to learn. The algorithm makes its own decisions regarding the operations to be performed to accomplish the task in question. This technique makes it possible to carry out much more complex tasks than a conventional algorithm. Andrew Ng, of Stanford University, defines machine learning as follows: “the science of getting computers to act without being explicitly programmed”. This encompasses the design, analysis, development and implementation of methods enabling a machine to operate via a systematic process, and to accomplish difficult tasks. Artificial intelligence grounded in machine learning therefore concerns algorithms which have specifically been designed so that their behaviour can evolve over time, based on their input data.

Deep learning is a subclass of machine learning, and forms the cornerstone of recent inroads made in the latter³. **A distinction is drawn between supervised machine learning⁴** (input data labelled by humans is given to an algorithm, which then defines the rules based on examples which are validated cases) **and unsupervised learning⁵** (unlabelled input data is given to an algorithm, which carries out its own classification and is free to produce its own output when presented with a pattern or variable). Supervised learning requires supervisors to teach the machine the output it must produce, i.e. they must “train” it. Such supervisors often accomplish a multitude of very simple tasks in practice. Platforms like Amazon’s Mechanical Turk are examples of places that recruit these thousands of “micro-workers” (sociologist Antonio Casilli) who, for example, label the vast amounts of photographs used to train an image recognition program. Google’s captcha system reCAPTCHA is another large-scale example of humans being employed to train machines. These machine learning algorithms are being embraced across a growing number of sectors, from road traffic prediction to medical imaging analysis.

³ This is a set of machine learning methods that attempt to model high-level abstractions in data using structured architectures of different nonlinear transformations. It adopts a similar approach to the functioning of neurons, hence why you often hear talk of “neural networks”.

⁴ A credit scoring algorithm will apply this technique: all of the customers’ and their loans’ known characteristics are input, and customers who have not repaid their loan are indicated; the algorithm will then be able to provide a score of the risk that future customers may not repay their loan.

⁵ An algorithm for detecting fraud typologies will employ this technique: an algorithm is given all of the data bearing on demonstrated fraud, and will be able to infer similarities between them so as to produce fraud typologies. Unsupervised learning can also be harnessed to identify the word sequences of different speakers on the waveband of a radio programme.



Example of image recognition

Image recognition is a good example for understanding the difference between conventional algorithms and machine learning algorithms (the latter often being labelled as AI today). Let's imagine that we want a machine to be able to recognise tigers. If we were to go about this by means of a classical algorithm, we would have to be able to explicitly describe in programming language all of the intellectual operations that we carry out when we realise we are looking at a tiger rather than another animal, even one such as a lion or cat for example.

Telling a tiger apart from a cat is easy, even for a small child. But breaking this process down to specify all of the steps necessary to recognise a tiger (in other words, providing the algorithm for this) is, if not impossible, then certainly a hugely daunting and time-consuming challenge. This is where the technique of machine learning comes in. For this involves providing the machine with a large amount of examples, in this instance scores of photographs of tigers, as well as photographs of other animals. From this dataset, the machine will learn to recognise tigers. By comparing the thousands of photographs input, it will work out, entirely on its own, the criteria it will use to recognise tigers in any photographs it may subsequently receive.

This is "supervised learning": humans supply the machine with thousands of photographs they have previously labelled as showing tigers, along with others that have explicitly been labelled as not showing tigers.

The example of image recognition (see inset above) gives an idea of how artificial intelligence is paving the way for the automation of tasks that are incomparably more complex than those handled by conventional algorithms. Unlike deterministic algorithms, AI taps into the data it receives to develop, with no outside help, the models it will apply to understand the situations with which it is presented. This is why it holds such promise today in sectors that generate huge volumes of data, such as meteorology.

There are already countless examples of AI in use – in many other fields than just form recognition. Accordingly, the classification of spam from among incoming messages on Gmail is a simple, typical example of AI in practice.

Google gathers a sizeable and constantly updated base from the spam reported by its users. The system then uses such information to learn how to determine what characterises spam. It can then decide itself which messages to filter. Artificial intelligence is at work in Google's machine translation service too. The company also claims to have used machine learning to analyse how the cooling system of its data centres works. Automation of this analytical function has reportedly enabled a 40% reduction in the amount of energy required to cool these facilities.



DID YOU KNOW?

A company like Airbus is already putting artificial intelligence into practice today for the purposes of form recognition. AI can allow to recognise the different vessels in an aerial shot of a maritime zone. What is the goal behind? Comparing the location of the craft thus identified with the signals sent by the beacons, so as to detect which ships are in distress or are looking to elude maritime surveillance for example. The interest lies in the swiftness of an operation which, if it is not automated, requires a great deal of time and resources. The progress made by such techniques in recent years is such that machines are now more reliable than humans in identifying ships that are sometimes difficult to tell apart from clouds.


Use of AI in industry is nothing new: it particularly gained traction in the 1980s, when it optimised the operation of nuclear power plant tank emptying, automating the computing and improving its reliability at the same time, by enabling substantial savings to be made thanks to shorter facility downtimes during maintenance.

Chatbots and voice assistants (Siri, Google Assistant or Alexa among them) are another fast-developing branch of artificial intelligence: they are capable of supplying information and answering standard questions, for example.


From these applications, we can see how **machine learning strictly speaking constitutes a disruption from conventional algorithms**. **Machine learning algorithms** are a whole new class of algorithms: we are steadily progressing “from a programming world to a learning world” (Jean-Philippe Desbiolles, public debate launch, CNIL, 23 January 2017). Classical algorithms are deterministic, their operating criteria are clearly defined by the people wishing to run them. Machine learning algorithms, on the other hand, are called probabilistic. Although they represent a much more powerful technology than the former, their output is always changing depending on the learning basis they were given, which itself changes in step with their use. Going back to the example of the tiger (see inset), it is possible that a form of artificial intelligence having been trained with a basis that only comprises one species of tiger may not be capable of recognising a tiger belonging to another species. But it is also feasible to assume that this AI is quite capable of bettering its skills at recognising other species of tiger as it comes across ever more cases with traits common to two species.

Beyond these technical differences, an overall approach to algorithms and AI nevertheless remains pertinent. Deterministic and machine learning algorithms alike raise common questions. In both cases, the end goal of applications making use of these classes of algorithm consists of automating tasks that would otherwise be performed by humans, or even to delegate to these automated systems more or less complex decision-making. Once we move away from a purely technical approach to these subjects, to consider their consequences and social, ethical and even political implications, the issues raised largely overlap and justify a joint investigation.

On a final note, in many respects algorithms and artificial intelligence also overlap with what is referred to in a generally vague manner as Big Data. This not only encompasses immense quantities of diverse data, but also the techniques for processing them, for getting them to make sense, for pinpointing unexpected correlations in them, and even for bestowing a predictive capacity on them. Similarly, artificial intelligence is inextricably bound up with the immense amounts of data required to train it and which, in turn, it can process.



**The algorithm without data
is blind. Data without
algorithms is dumb**



Steer the discussions towards the most crucial impacts and applications of algorithms today

In one sense, algorithms tie in with computer science and, more generally, with everything we tend to group under the term “digital”.

With such a potentially vast subject as this, it is both necessary and entirely justified to limit the scope of our discussions to the algorithms which are currently raising the most pressing ethical and societal questions. **Indeed, an ethical discussion on AI systems and algorithms will only be meaningful if it also devotes thought to the implications these have in social, human and professional contexts.**

In the pages that follow, we will thus be focusing solely on those uses of artificial intelligence grounded in machine learning – the most widely discussed today even if, strictly speaking, they do not account for the whole of this domain⁶.

Moreover, strong AI (artificial general intelligence) has not been included in our analysis. This refers to systems that are capable of becoming completely autonomous, to the point that they even turn against humans. This vision is often fuelled by an apocalyptic mindset inspired by the movies in the wake of sometimes much older myths (Frankenstein, etc.). It is often connected with questions over the level of consciousness that such a machine could attain (in keeping with the theme of technological singularity). Strong AI is promoted by stances adopted by high-profile digital leaders, of the likes of Elon Musk or Stephen Hawking. Lastly, the promotion of the theme of the “singularity” by transhumanist circles extending their influence from Silicon Valley is lending further credence to the claims that machines will soon surpass humans. And yet the foremost researchers and experts in computer science, France’s Jean-Gabriel Ganascia among them, are sceptical of such claims. **Some (including the latter) even lambast this hypothesis of a looming strong AI as a**

means of evading more serious issues – be they ethical or quite simply legal – that the tangible achievements made with weak AI, and its rising use, are already or will very shortly be raising.

Strictly speaking, and by taking the terms quite literally, it would have been possible to include within the scope of our discussions on algorithms the questions to do with encryption, insofar as this technology relies on the use of algorithms. The same line of thinking could have led us to consider the “blockchain” as an integral part of the subject. But, as before, it seemed preferable to adopt a pragmatic stance, guided by public perception of which algorithms and their applications are raising the most problems and questions today. In other words, we have chosen to limit our thought process to those algorithms, in all their immense diversity in this digital age, which are today raising issues most likely to directly concern the general public and public and private decision-makers alike.

From this point of view, although **recommendation algorithms** technically only account for a fraction of the different types of algorithms out there, they are an important part of the question. These algorithms are used to establish predictive models from a large amount of data and to apply them in real-time to specific cases. They develop predictions on behaviours or preferences which make it possible to anticipate consumers’ needs. An individual can be steered towards a choice that has been deemed most appropriate in his or her regard. Recommendation algorithms can be used to make suggestions of restaurants on a search engine, for example.

If we take this approach further, we can thus list the main **functions and applications of algorithms** that are likely to be controversial, and which this discussion addresses:

⁶ The two overarching approaches to AI are, on the one hand, the symbolic and cognitive science-inspired approach and, on the other, the neuroscience-inspired and connectionist approach (machine learning, neural networks, etc.). Expert systems gained considerable ground in the 1980s. The main advances made recently all concern machine learning.

- Producing knowledge;
- Matching supply and demand, allocating resources (passengers and taxi drivers, parents and childcare places, students and university places, etc.);
- Recommending a product, a special offer in a personalised way;
- Assisting with decision-making;
- Predicting, anticipating (natural phenomena, offences or the onset of an illness for example).

These significant functions are enabled by algorithms' capacity to filter information and model phenomena by identifying patterns in massive datasets and thus to profile individuals⁷.

Generally speaking, **the high public profile of algorithms and the questions they raise today cannot be considered in isolation from the unprecedented volumes of data now available across all sectors, which need to be sorted to harness their full potential.** The digitalisation of our society in all its forms – electronic transactions and services, revolution of sensors, the Internet of Things, surge in smartphone use, broad-scale roll-out of open data policies and so on – is behind this phenomenon which, whilst representing an invaluable resource today, also poses a challenge. **If we need recommendations, it is because of the sheer amount of information now available; if it is possible to profile, it is because mere segmentation by pre-determined categories is no longer sufficient in light of the large amount of data now collected on data**

subjects. The quality and relevance of the data chosen as input for algorithms are further key considerations for any discussion in their regard.

The idea of **autonomisation** must also be mentioned, if we are to correctly gauge the issues raised by algorithms today. For these questions also stem from the fact that algorithms make it possible to delegate tasks, previously accomplished by humans, to automated systems that are becoming increasingly "autonomous". The delegation of tasks, and decisions even, to conventional algorithms does not in any way imply, however, that algorithms themselves are free from human control. Human intervention is well and truly present in the use of algorithms, through the algorithm's configuration, through the choice and weighting of both data and the criteria to be taken into account to obtain the desired output. For example, although humans are no longer directly involved in the suggestion of restaurant recommendations on platforms using algorithms, developers still have a fundamental role to play. They determine, in particular, which importance to give to information such as places where restaurants are located, their rating by other users or their supposed match (here again based on criteria to be defined) with the user's profile.

With the rise of machine learning, we are one step closer in this autonomisation momentum, since the machine "itself" is writing the instructions it performs and determining the parameters that will guide it in accomplishing a goal, the latter being nevertheless still defined by humans.

**Algorithms today cannot be considered
in isolation from the unprecedented volumes
of data now available across all sectors,
which need to be sorted to harness their full potential**

⁷ The General Data Protection Regulation defines profiling as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".

Use and promise across all sectors

Use of algorithms and artificial intelligence is growing across all sectors. Economic stakeholders are keen to promote the advantages and promise held by such tools. We will mention a few examples here⁸.

The most commonplace uses today particularly concern online search engines, road navigation apps, cultural content recommendation on platforms (of the Netflix or Amazon type) or social media and marketing for targeted advertising purposes or, increasingly, for political campaigning during elections.

Regarding healthcare, the use of algorithms is promoted for health surveillance (detection of epidemics or mental health risks). There is increasing talk of the promises of precision medicine where personalised therapeutic solutions are developed by cross-linking patient data with the datasets obtained from large-scale cohorts.

The State's governing powers are also concerned. Some actors for example claim to provide the legal occupations (a judge, for instance) with tools which would enable them, by processing case law data, to anticipate the outcome of a trial or fine-tune a judicial strategy. Meanwhile, police forces in France and abroad are beginning to use algorithmic tools to channel their resources towards a given area through data analysis.

The wide media coverage on the French university admissions platform "Admission Post-Bac" opened the general public's eyes to the use of algorithms for allocating university places to hundreds of thousands of pupils. Over and above the flow management, algorithms are challenging teaching practices via ever more advanced strategies to personalise education or via the detection of potential early school leavers.



SURVEY

Patchy knowledge of algorithm use*

The general public is well aware of the use of algorithms in targeted advertising for example (90% of respondents are wised up to this).

But they are less aware of the reliance on algorithms to assess "relationship compatibility" on dating applications (46% of respondents) or to establish a medical diagnosis (33%) for example.

* Survey carried out as part of the public debate by the rural-based family association "Familles rurales", among 1,076 of its members.

Finally, on the job market various stakeholders are currently working on developing solutions for assisting with recruitment (by matching supply with demand in particular) and managing human resources.

Without claiming to provide exhaustive coverage of a subject with countless applications, the table on the next page nevertheless gives an idea of the way in which artificial intelligence's and algorithms' main identified functions are in use across different sectors.

⁸ Industrial development of artificial intelligence is mainly being driven by two types of stakeholders. On the one hand, specialist providers of services and technology to large companies, such as IBM with Watson. On the other, the digital data giants (including the GAFA), which are investing heavily in AI and making it a central feature of their services (Google with Translate, image recognition or automatic speech recognition for example).

The main functions of algorithms and AI across different sectors

	Education	Justice	Health	Security	Work, HR	Culture	Other
Generating knowledge	Better identify learners' abilities	Reveal the different ways judgments are handed down between regions	Tap into the vast amount of scientific publications	Identify unsuspected links for solving gendarmerie-led investigations	Understand social phenomena in the workplace	Create cultural showpieces (painting, music)	Fine-tune an insurance company customer's risk profile
Matching	Allocate higher education places to candidates (APB)		Allocate patients for participation in a clinical trial		Match a list of applicants to a job vacancy		Match "compatible" profiles on dating apps, etc.
Predicting	Predict early school leaving	Predict the likelihood of a trial being successful and the potential amount of damages	Predict epidemics Pinpoint predispositions to certain diseases to prevent their onset	Detect at-risk profiles in counterterrorism strategy Predict future crimes and offences	Detect any employees who are likely to resign in the coming months	Create crowd-pleasers (Netflix)	
Recommending	Recommend personalised learning pathways to students	Recommend mediation solutions based on the profile of the individuals and similar cases in the past			Recommend career guidelines in line with individual profiles	Recommend books (Amazon), TV series (Netflix), etc.	Personalise political messages on social media
Assisting with decisions		Suggest to the judge the most fitting case-law solution for a given case	Suggest suitable therapeutic solutions to the doctor	Suggest hotspots for police forces to patrol			Help drivers to find the shortest route from A to B (GPS)

The ethical issues

Ethics, as a prefiguration of legal standards

P.24

Autonomous machines: a threat to free will and responsibility?

P.26

Bias, discrimination and exclusion

P.31

**Algorithmic profiling: personalisation
versus collective benefits**

P.34

**Preventing massive fires while enhancing AI:
seeking a new balance**

P.38

Quality, quantity, relevance: the challenges of data curated for AI

P.39

Human identity before the challenge of artificial intelligence

P.41

The ethical issues

Ethics, as a prefiguration of legal standards

The notion of ethics is often attributed different meanings, at times giving rise to a certain ambiguity. The definitions given in the dictionaries liken ethics to morals, in other words to standards which are not necessarily intended to come within the realm of law and which have to do with individual behaviour. For the ancient philosophers then, ethics is nothing more than the answer to this question: "what is a good life?" – i.e. guidelines for action which, first and foremost, concern the individual.

Fast forward to more recent times, the notion of ethics has evolved to refer to something alongside the law, used by stakeholders such as companies. In this instance, ethics is a set of standards laid down by the company, and by which it obliges itself to abide. These standards can go beyond the legal sphere. Often, their only purpose can be to restate – consciously or otherwise – legal standards. Some examples of the "ethical" use of customer data are sometimes merely another way of saying that the company is legally compliant.

A third meaning attributed to the notion of ethics – and arguably the most relevant within the context of this report – has emerged in the jargon of public institutions since the setup back in 1983 of the French governmental advisory council on bioethics issues (CCNE). In this context, **ethics comes across as a guiding process in legal matters, and the ethical standard as a prefiguration of the legal standard**. So the fact that the lawmaker asks an institution to engage in ethical thinking means that this will soon be followed by a corresponding legislative framework. The setup of the CCNE by the law shared a key common denominator with the creation of the Digital Republic Bill and its own provision for an ethical discussion assignment entrusted to the CNIL: a backdrop of swift technological progress and grave uncertainties over the attitude that the community should subsequently be adopting. On the one hand, the breakthroughs in biotechnology (the first test-tube baby in France was born in 1982), on the other what is being billed as a "digital revolution". Making an

SURVEY

The public show a measure of distrust in algorithms and AI*

The three most commonly held fears are **loss of human control** (63% of members), **normativity and restriction through the standardisation of recruitment** (56%), and the **disproportionate collection of personal data** (50%).

In recruitment and human resources, a few opportunities have been highlighted, including the possibility of examining all applications on the basis of identical criteria (52%). That said, for **72% of respondents, the possibility of being recruited by algorithms**, on the basis of an analysis of their profile and compatibility with a specific job, is perceived as a **threat**. 71% of them thus state that the definition of an ethics charter on algorithm use is a genuine priority.

72 %
of respondents perceive the possibility of being recruited by algorithms as a threat

* Survey carried out as part of the public debate by the CFE-CGC, a trade union for executives in France, among 1,263 of its members (primarily from the "Metallurgy" and "Banking & Finance" federations).

ethical thought process part of the law is therefore a way of laying the necessary groundwork for holding collective discussions on a social pact. The latter indeed see some of its founding principles (fundamental freedoms, equality between citizens, human dignity) being called into question when the technological shift moves the boundary between the possible and the impossible, and calls for the distinction between the desirable and the undesirable to be redrawn.

For this first discussion, the CNIL decided to call on stakeholders who were keen to express their views on subjects to do with algorithms and artificial intelligence. The ethical issues we have chosen to focus on are therefore the very ones that these stakeholders raised themselves. Logically, most of them are already a firm feature of our everyday lives (even if they are likely to gain even greater ground in the years to come). On the other hand, more forward-looking challenges, associated with progress that is still hypothetical (transhumanism, human-machine

hybridisation, etc.), were not uppermost in the minds of the partners involved and, as such, are not greatly delved into in this report.

The technological shift is moving the boundary between the possible and the impossible and calls for the distinction between the desirable and the undesirable to be redrawn



WHAT THE PUBLIC THINK

Participants in the public consultation organised by the CNIL, in Montpellier on 14 October 2017, shared their thoughts on the ethical matters raised by algorithms and artificial intelligence (see “An innovative approach to crafting a collective and pluralist ethical thought process”): the issues they bring to light chime largely with those identified throughout the public debate.

The public seems most anxious about the new decision-making processes being adopted and the watering-down of liability as a result of algorithms. The **potential “loss of competence” on the part of doctors or employers** who would come to rely heavily on algorithms has been highlighted. The detrimental consequences mentioned include automated “management of uncertainties” that is not deemed as effective as what humans are capable of doing; an inability to “manage exceptions” and the “loss of a sense of humanity” (this was particularly emphasised with regard to the lack of any appeal options on the “Admission Post-Bac” (APB) portal).

Reliance on sometimes autonomous IT systems to make decisions is stoking fears that **liability** in the event of error is “not clear” – a concern voiced particularly about the medical sector. On the subject of APB, some members of the public criticise the lack of transparency, which explains why the algorithm serves “as a scapegoat, creating a distance between those who make political choices and those who complain about these choices”. The problem of information personalisation on social media and its collective effects, touched on in the context of the US presidential elections, also heightens their fear that “no one is really accountable for supervising the Internet anymore”.

The danger of filter bubbles, though brought up less often, was nevertheless mentioned by several participants in the “human resources” and “digital platforms” workshops. The public also spoke of the risk of recruitment “**standardisation**” and the subsequent streamlining of a sector that should not be so, as well as the risk of being boxed online “into a profile that might hamper our personal growth”.

Lastly, the subjects of **bias, discrimination and exclusion** warrant particular vigilance in the participants’ view, irrespective of whether the bias in question is intentional (with respect to recruitment, there are fears that an algorithm could be coded “according to employers’ objectives, at employees’ expense”) or unintentional (the algorithmic tool is a source of concern in terms of the errors it could generate).

Autonomous machines: a threat to free will and responsibility?

Beyond the sheer number of practical applications and uses to which they can be put, **the purpose of both algorithms and artificial intelligence alike is to automatically accomplish a task or operation involving a form of "intelligence", which would otherwise be carried out directly by humans.** In other words, this entails humans **delegating tasks to automatic systems**⁹.

The case of the "Admission Post-Bac" (APB) university admissions online portal is a good example. This software allocates higher education places to high school graduates. It could be considered as doing nothing more than applying a set of instructions and criteria that could just as well be carried out by civil servants. The key interest in using an algorithm in this instance is the gain in productivity brought about by delegating to a machine a task that would take up a great deal of time and resources to a human. Another interest is to guarantee the uniform and impartial application of rules defined beforehand for the allocation of available university places. Indeed, the implementation of these rules by a complex administrative chain of command is much more likely to give rise to forms of arbitrary decision-making or even simply different interpretations depending on the staff applying them. In this regard, educational policy specialist Roger-François Gauthier points out that APB at least has the merit of putting paid to a "Mafia-like" system which practised preferential treatment¹⁰.

APB is an example of a classic deterministic algorithm. Artificial intelligence can however also be harnessed to accomplish tasks that might otherwise prove too costly in terms of human resources. For example, **form recognition is used to identify, in real time, vessels on satellite images** over vast maritime surface areas. A simple software program can thus perform round-the-clock monitoring of immense surface areas which would otherwise require shift work on the part of several specialists.

In the future, it could be possible, at least technically speaking – and this is already in progress in the United States – to assign algorithms the task of determining the threat posed by an inmate, and therefore the opportunity of granting remission. The next step of what some refer to as "predictive justice" would involve entrusting systems with the task of making decisions based on a cross-analysis of the data pertaining to a certain case, with case-law data.

Delegating tasks to algorithms: contrasting situations

What immediately becomes clear is that **concern over the potential ethical and social implications of automated systems varies** depending on the tasks being delegated to the latter and the very conditions shaping this delegation.

Accordingly, a further step can be taken in distinguishing which cases the discussion should concentrate on. We can think of a **typology of task delegation to automated systems**, grounded in two criteria: **the impact on individuals of the task being delegated, and the type of system this task is being delegated to.**

The first criterion concerns the type of impact and/or scale of the task delegated to the automated system. It might be a routine, mechanical and fairly innocuous task (such as sorting a series of computer files into alphabetical order). On the other hand, this task might lose its trivial nature and prove to be hugely complex. It could, above all, assume aspects of a decision and take on vital importance for an individual or group. It is such a case when it involves establishing guidance for a medical diagnosis. Between these two extremes lies a vast spectrum of contrasting situations. These include the two aforementioned examples, as well as that of the autonomous vehicle: this and the case of APB have more in common with the case of the automated medical diagnosis than examples at the other end of the spectrum.

The second criterion would have to do with whether the automated system relies on a conventional algorithm or a machine learning algorithm. **In other words, it is about the degree of autonomy of a system:** is it able to establish its own operating criteria? What is also at stake here is the extent of the system's ability to produce a satisfactory explanation for the output it produces.

This typology shows the wide diversity of situations to be covered in a discussion on the ethical and social issues of algorithms and artificial intelligence. It above all sheds light on the sheer range of the spectrum of serious or less serious matters in the use of such or such an algorithm.

⁹ Strictly speaking, remember that it is not generally so much the use of algorithms in itself that is new, as its execution in the form of a computer program.

¹⁰ Public debate launch, CNIL, 23 January 2017.

Delegating critical decisions to algorithms: absolving ourselves of responsibility?

Some crucial decisions (medical diagnoses, court judgments, decisions to open fire during armed conflict) could be and sometimes are already (abroad in particular) delegated to automated systems. Such decisions are often already clearly identified by legal tradition in France. Only a doctor is thus qualified to establish a diagnosis: if not, it would be unlawful medical practice. The same applies for a judge's decision which, legally, has no place being delegated to an automated system. With this in mind, algorithms endorse the role of decision "support" so as to increase humans.

Such laws do not, however, solve all the problems raised by the delegation of decisions. **How can it be ensured that the predictions and recommendations provided by algorithms remain nothing more than supports to human decision-making and action, instead of leading to humans no longer being held to account, meaning an overall loss of free will ?**

In medicine, the quality of decision-making can be assessed (or at least quantified) more easily. We might logically ask ourselves how much scope of autonomy would doctors keep despite the recommendation (for a diagnosis or treatment) that had been supplied by a cutting-edge decision support system. Indeed, artificial intelligence is hailed as having the potential to diagnose certain cancers or analyse X-rays with more accuracy than humans. Were this to prove correct, it could therefore become risky for a doctor to make a different diagnosis or therapeutic choice from the one recommended by the machine. In this case, the machine would become the official decision-maker. This kind of scenario raises the question of liability. Should this be transferred to the machine itself, which would then need to be granted of a legal personality? Or to its designers? Or should the doctor still be held accountable? But in that case, although this certainly seems to resolve the legal problem, does it not still ultimately lead to a de facto absolution of responsibility, to the development of a sense of irresponsibility?



FOCUS

Ethics and predictive policing

Algorithmic software is being actively developed in the quest for a form of crime prediction in space and time. Its goal is to predict where and when crimes are most likely to occur, in order to provide patrolling guidance to law enforcement. In the United States, **PredPol** is grounded in models inspired by seismology to assess how intense a risk is in a given location at a given time. The startup thus claims to factor in the "contagious" dimension of the spread of crime in space and time.

This predictive potential has nevertheless shown its limits, on the one hand because contagion has a much smaller impact on crime detection than the after-shocks of an earthquake and, on the other, because the structure of criminality varies from one year to the next. And yet, this does nothing to dampen the appeal of such systems which help to **"handle, on the basis of the management criteria, the daily patrolling of law enforcement officers"**. In practice, *"the predictive box stays red on the map as long as the police has not yet patrolled that location; on their arrival it turns blue and, when the police officer has spent the sufficient, optimum amount of time there as calculated according to the resources available, it finally turns green"*¹¹.

This raises a major concern: what about the risk of the machine's recommendations being taken as an absolute truth, which does not require any discussion as to the practical consequences? Insofar as the algorithm uses data from victims' statements, one such consequence is beefed-up police presence in zones where the population has reported crime in a more sustained fashion. This implies the exclusion of the public protection provision for certain populations (those who do not report crime as often). It would also not be far-fetched to imagine this type of algorithm focusing police attention on certain types of offence to the detriment of others.

The bottom line is that a critical assessment of this type of tool is paramount. What about the capacity to determine the effectiveness of such models? Whether or not a crime is detected by a patrol taking its cue from the system, the outcome could easily (but falsely) be interpreted as a sign of the tool's effectiveness.

¹¹ Bilel Benbouzid, "From situational crime prevention to predictive policing: Sociology of an ignored controversy"

The case of medicine is particularly critical. Not only on account of the impact that decisions and recommendations have on individuals, but also because, in this instance, machine learning algorithms are at play. This implies that the underlying processes at work in AI systems are potentially as incomprehensible to the people to whom they are proposed as they are to their designers themselves. Incidentally, the CNIL-led public debate brought a controversy to light in this regard, concerning IBM's Watson platform. IBM's policy stresses that Watson is trained via "supervised learning". In other words, the system is guided, step-by-step, in its learning. This should mean its process can be monitored, as opposed to unsupervised learning, in which the machine has full autonomy in determining its operating criteria. IBM also claims to check what the systems have been doing, before any decision to retain a certain type of learning. But experts researching this subject who have spoken out during the various organised debates (not least by Allistene's research committee on ethics, CERNA) have insisted time and again that such statements are erroneous. Based on current research, the "output" produced by the most recent machine learning algorithms is not explainable, explainable AI being a concept on which research is ongoing. They also point out that it is very difficult to audit a machine learning system in practice.

We might therefore wonder whether algorithms and artificial intelligence are not undermining traditional figures of authority – decision-makers, leaders – and perhaps even the very authority of the rule of law. This trend is sometimes actively supported. Some, Tim O'Reilly among them, are already considering the advent of "algorithmic regulation"¹² where city "governance" would be entrusted to algorithms. Sites, infrastructure and citizens would constantly communicate data processed thanks to smart sensors. This could streamline and optimise community life according to so-called "natural" laws. These laws would stem from the way things really are, an "immanent normativity" as explained by Thomas Berns and Antoinette Rouvroy¹³. This undoubtedly sheds light on the temptation to turn away from human normativity towards an "algorithmic normativity" driven by market incentives. These discourses extol the supposed "objectivity" of automated systems (in comparison with human judgment which is always fallible). Users are thus increasingly willing to take the output produced by a machine to be an unquestionable truth – when it is actually determined throughout by human choices (criteria, types of data fed into the system)¹⁴.

The impact algorithms have on decision-making could also come in a different form. In the report it submitted to the CNIL, the Conseil National des Barreaux, the national institution that represents all practising lawyers in France, highlighted that **"care must be taken to ensure that the obsession for effectiveness and predictability behind the use of algorithms does not lead to us designing legal rules and categories no longer on the grounds of our ideal of justice, but so that they are more readily 'codable'"**.

It is quite possible that this ongoing shift towards forms of "algorithmic regulation" holds a certain appeal for decision-makers themselves. Delegating decisions to a machine – which we credit with being neutral, impartial and infallible – may be a way to remove responsibility from oneself, to wash one's hands of the need to be accountable for one's choices. **Autonomous lethal weapons (killer robots)** which could, themselves, make the decision to kill on the battlefield or for law enforcement purposes, throw this question into particularly sharp relief. Should the act of killing – even when considered lawful in a situation of international conflict and when the enemy is armed – remain under the control and direct responsibility of humans? Should its difficulty and potentially traumatic implications for the person carrying it out be deemed a necessary guarantee for avoiding abusive practices?

These considerations do not just concern situations where tasks or decisions are delegated to a machine learning algorithm. The traditional, deterministic, algorithm is also concerned. The debates on APB's algorithm have provided a good example of this in practice. APB gives us insight into how **it is possible for such a process to take root where societal choices are depoliticised and neutralised – even though a public discussion is warranted in this regard.** The controversy focused on the algorithm itself, particularly in the wake of the revelation of the drawing of lots it ended up carrying out for certain applicants in over-subscribed courses. But algorithms are only ever the reflection of political choices, society choices. In this instance, the choice to draw lots to allocate places in over-subscribed courses is the result of a political choice. Two possible alternatives to which could have been, put simply, selection upon admission to university, or investment to increase the number of available places in the courses in question, so as it matches demand. In other words, "code is law", as famously asserted by Lawrence Lessig.

¹² Tim O'Reilly, "Open data and algorithmic regulation", ed. Brett Goldstein, *Beyond Transparency: Open Data and the Future of Civic Innovation*, San Francisco, Code for America, 2013, pp. 289-301.

¹³ Rouvroy Antoinette, Berns Thomas, "Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation ?", *Réseaux*, 2013/1 (issue no. 177), p. 163-196.

¹⁴ The so-called "machine objectivity" is actually only diluted, unassumed subjectivity in this respect.

Indeed, "neutral" is not a description we can ever ascribe to an algorithm (understood in the broad sense as the socio-technical system into which it fits). It inevitably incorporates bias – be this social, political, ethical or moral – and usually meets purposes which include a commercial dimension for its designer. A commonly cited example is the choice that the algorithm of a driverless car might be forced to make between killing its occupant or a pedestrian on the road. It illustrates how reliance on technology does more than raise certain moral dilemmas; it above all moves them to a different stage: a quandary settled in real time by a person directly involved is replaced by a choice, made by others, elsewhere, well beforehand¹⁵.

From the intentional purpose of setting up APB (enhanced administrative effectiveness and fairer harmonisation of the allocation of higher education places), we ended up with the evasion of society choices hidden in the configuration of the system but masked by the assumed impartiality of the algorithm. Those responsible for implementing the algorithm that takes decisions must therefore look for ways to counter this type of effect (by informing the target audience for example). They must systematically take care neither to exploit it by hiding behind the machine, nor to allow it insofar as it tends to keep conflicts or legitimate debates at bay.

Moreover, it is likely that such behaviours will lead to a feeling of inhumanity in the individuals concerned. It would risk turning into distrust, especially where there is no pos-

sibility of contacting the managing body and talking in a bid to "find solutions or just to be heard", as highlighted by the French mediator of National Education¹⁶.

For a deterministic algorithm such as the one discussed here, the watering-down of responsibility is just an illusion. The crucial decisions and choices have simply been moved to the configuration stage of the algorithm.

Is this tantamount to saying that those who control the computer code are becoming the real decision-makers, and that there is a risk of all the power becoming concentrated in the hands of a "small caste of scribes" (Antoine Garapon, debate launch, 23 January 2017)? This is at least not the picture we get from the APB example at any rate. Following the opening of the source code of the authorities' algorithms as required by the Digital Republic Bill, APB's algorithm was examined by the Etalab mission (which works in France on data sharing in the public sector). What became clear was that its developers had taken care to document within it the origin of each change in configuration of the algorithm, in this case each instruction they had received from the government authorities. In a nutshell, traceability in terms of accountability had been organised by the developers of APB themselves. For all that, this example should not mask the fact that algorithms tend to bring the decision-making forward to the technical stages of a system's design (configuration, development and coding). The system then only results in the automatic and flawless implementation of the choices made initially. Antoine Garapon's aforementioned concern is therefore well-founded and demands answers. **It is essential that these design stages do not become so independent that they are where the decisions are made.**

The question of where accountability and decision-making can be set up is to be approached in a slightly different way when dealing with machine learning systems. In this case we would do better to think more in terms of chain of accountability, from the system designer right through to its user, via the person who will be feeding the training data into this system. The latter will operate differently depending on such input data. On this subject we could mention the Microsoft's chatbot Tay. It was shut down a mere 24 hours later its release when, learning from the posts of social media users, it had begun to tweet racist and sexist comments of its own. Needless to say, working out the precise share of responsibility between these different links of the chain is a laborious task. From this, **should we base the use of artificial intelligence on the condition that this liability can be attributed with absolute certainty?** We already know that artificial intelligence can outperform

Algorithms and artificial intelligence are in some ways undermining traditional figures of authority, decision-makers, leaders, and perhaps even the very authority of the rule of law

¹⁵ On this subject, see MIT's commendable website, which provides a practical illustration of these dilemmas : <http://moralmachine.mit.edu/>
¹⁶ Le Monde, 29 juin 2016 : "Le médiateur de l'Education Nationale dénonce la solitude des familles face à APB".

humans when it comes to certain tasks, without properly understanding how these systems work nor, as a result, any errors they might commit. Rand Hindi thus explains that "AI makes fewer mistakes than humans, but makes mistakes where humans would not have done so. This is what happened with Tesla's driverless car accident, which never would have happened with a human". Should we consider, then, bestowing these systems with a legal personality? Or hold users themselves responsible (so, in the medical sector for example, this means the patients)?

We would be wise not to over-stress the specificity of the machine learning case for all that. Imagine a form of artificial intelligence tasked with allocating patients within a hospital's departments and with determining the end of their hospital stay in the most "effective" way possible. There is bound to be some opacity inherent in such a machine learning system. Be that as it may, the objectives set in its regard, as well as their weighting (guaranteeing as many recoveries over the long term as possible, minimising the rate of rehospitalisation in the short term, aiming for short hospital stays, etc.), would still be choices made explicitly by humans.

A question of scale: the massive delegation of non-critical decisions

Should ethical thinking on algorithms and artificial intelligence be limited to crucial decisions, sectors where the impact on humans is undeniable, such as medicine, justice, educational guidance, and even the automotive sector with its implications in terms of safety? **Should attention not also be paid to those algorithms to which we are gradually delegating more and more apparently innocuous decisions but which, taken together, form the substance of our everyday lives?**

Simply on account of their ability to operate in a repeated manner, over long timeframes and above all at vast scales, algorithms can have a substantial impact on individuals or societies as a whole. Take the criteria underpinning the operation of a basic navigation app for example. It can significantly alter the very form of the city and urban life as a whole, when used by a large number of drivers, all placing their implicit trust in this app to map out the itineraries they should take, the repercussions on urban traffic, the spread of pollution.

The CNIL's Digital Innovation Laboratory (LINC) puts it this way: "Aside from the question of the collection of personal data, there is also the issue of the public stakeholder's loss of control over the planning of public space, the management of flows and, beyond that, the very notion of public service and general interest. The individual interests of a Waze app's customers, when taken together, can sometimes be at odds with a local authority's public policies"¹⁷.

In her book *Weapons of Math Destruction*¹⁸, Cathy O'Neil provides a particularly meaningful example. She imagines that she could draw up the rules she implicitly follows to plan her children's meals (diversity, green vegetables but not too many so the kids don't protest too much, easing of the rules on Sundays and special occasions, etc.). A program running such an algorithm would be fine as long as it were only used to automatically generate a meal planner for a limited number of people. But the specificity of algorithms executed by computer programs is their scale of application. A program of this kind, when used as is by millions of people, would inevitably have powerful and potentially destabilising effects on major social and economic balances (certain foods would become more expensive while the production of others would collapse, standardisation of production and impact on agri-food occupations for example). **This very specific aspect of computer algorithms deployed in the Internet age we are currently living in, and which is laid bare by the author, is an altogether new challenge to be addressed: their scale of deployment.** And anyone deploying algorithms that are likely to be used on a large scale should be urged to bear it in mind.

Algorithms are compressing time frames

One of the defining features of the way an algorithm works is its immediacy and simplicity, or at least its uniformity and inexorable character. AI algorithms are capable of accomplishing a task almost immediately (simply in the time it takes for the machine to compute it). They are also capable of accomplishing this same task on a very large scale in spatial terms, and everywhere identically at that. In this respect, they can hold considerable appeal for the authorities or businesses committed to an effective, rational and standard way of working.

¹⁷ CNIL (LINC), La Plateforme d'une ville. Les données personnelles au cœur de la fabrique de la smart city, Cahier IP n°5, octobre 2017, p. 20.

¹⁸ Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

But this characteristic of algorithms is also potentially problematic: **compressing the duration and spatial dimension of a process delegated to a machine can also undermine the quality of the action.** Practical examples of algorithms being used by the authorities as well as the example of predictive justice give a clearer idea of this ambivalence, between the optimising and diminishing of processes stripped of their spatial dimension.

What this means is that using an algorithm along the lines of the APB platform's may certainly, from the authorities' point of view, be seen to guarantee a form of simplicity and consistency in the application of rules. A complex chain of administration involving many people may instead give rise to differences of interpretation and arbitrary choices. But might it not also be possible to consider something which, at first glance comes across as a lack of effectiveness or a sign of an somehow erratic process, as something else. Couldn't it be instead an invaluable source of information for decision-makers, through the feedback and questions reported by those who are in charge of applying the rules and are in a position to observe on the ground how it works, the limits?

Similarly, at the symposium on predictive justice organised on 19 May 2017 by the Lille Bar, Law Department of Université catholique de Lille and the Douai Court of Appeal, certain participants stressed that "knowledge of judgments given by the other neighbouring jurisdictions or by the other magistrates would contribute towards a certain consistency and prevent that the outcome of a dispute depends on knowing whether it is heard in a city or another". The idea here lies in the ability of algorithms to handle large amounts of case-law data that have been made "open" and to reveal disparities in the application of

the law across different jurisdictions. The identification of such disparities, of which the judge is not aware, would facilitate consistent application of the law nationwide. And yet, are we absolutely certain that, within certain limits, forms of regional disparity do not, in fact, reflect a responsible exercising of caution on the part of the judge? Its intelligent and detailed adaptation to social realities that can vary from one place to the next? Is it a way of allowing some leeway in how the law is applied, apart from its rigid and automatic application?

The same line of thinking could be applied to the idea of a predictive justice. A judgment handed down by artificial intelligence (the idea taken to the extreme) would circumvent the benefit of collective deliberation; meaning what can be gained from individuals working towards a common goal and comparing their points of view. **The deliberations carried out by juries and magistrates do not solely involve trotting out pre-existing arguments in the way software "executes" a program. The time such deliberations take is not simply a minor detail, a resource whose cost should be kept to a minimum: on the contrary it is of the utmost importance.** For it enables juries to take on board new insights over the course of hearing the different arguments, and to change opinion, as shows more clearly than any demonstration the film by Sidney Lumet, *Twelve Angry Men*.

At the end of the day, it seems advisable to draw the attention of users of algorithms and artificial intelligence to the need to heed not just the advantages but also any disadvantages of these technologies (and their potentially ambivalent nature), and to think about ways to overcome these.

Bias, discrimination and exclusion

The tendency of algorithms and artificial intelligence to generate bias which can, in turn, spawn or reinforce discrimination has raised significant concerns and questions. It is essential to highlight this point since these technical systems can also sometimes fuel a belief in their objectiveness (which is all the more valuable since it is often lacking in humans). And yet all algorithms are biased in a sense, insofar as they are always the reflection – through their configuration and operating criteria, or through their training input data – of a set of societal

choices and values. The debate raging on bias and discrimination appears as a magnifying glass which shows up all of the problems associated with this key characteristic.

Several recent controversies have illustrated this type of bias in a particularly shocking way. In 2015, Google Photos, a face recognition software, thus caused an uproar when two young African-Americans realised that one of their photos had been tagged as "Gorillas".

This glitch can be explained by the type of data with which the algorithm was trained to recognise people. In this instance, it is likely that it was primarily – if not exclusively – trained with photographs of white people (other examples of racist bias exist in face recognition software to the detriment of “Asian” people). As a result, the algorithm considered that a black person had more points in common with the object “gorilla” that it had been trained to recognise, than with the object “human”.

Note, also, that deliberate malicious acts on the part of people involved in training this type of algorithm are not unknown. This was the case with the chatbot Tay, developed by Microsoft, which began to post racist and sexist tweets (after just a few hours of working) based on the posts of other Internet users.

Algorithms can also exhibit gender bias. In 2015, three researchers at Carnegie Mellon University and the



Algorithms to prevent recidivism?

Predictive justice applications are being subjected to particularly close public scrutiny as regards their Management Profile for Alternative Sanction) tool designed to come up with a **recidivism risk score** for prisoners or defendants on trial. Although data analysis tools had been in use in **US courts since the 1970s**, automated score calculations for making decisions bearing on conditional release is something new.

In other words, social workers using COMPAS have access to an interface where, together with the defendant, they can answer questions such as “What does the defendant think of the police?”, “What are the defendant’s friends like?”, “Have some of them already been convicted?”¹⁹. A risk score is then calculated and added to the defendant’s file.

The ProPublica website accused Northpointe, the company marketing COMPAS, of producing **biased and racist scores**²⁰. This finding is based on the comparison of the released prisoners’ **recidivism risk scores to their actual recidivism rates** in the two years after they were scored. The rate of false positives (i.e. a high score but with no subsequent reoffence observed) turned out to be considerably higher for former prisoners of African-American origin than for white people.

International Computer Science Instituteshowed how AdSense, Google’s advertising program, generated bias against women. Using a testing tool called Adfisher, they created 17,000 profiles and simulated their browsing on the Web to conduct a series of experiments. What they found was that **women were much less likely to be displayed job ads for highly-paid positions than men, for similar levels of qualifications and experience**. It emerged that a limited number of women received online ads for a job earning more than \$200,000 a year. Far from being a one-off, “targeted advertising like Google’s is so ubiquitous that the information shown to people could have tangible effects on the decisions they make”, says Anupam Datta, co-author of the study.

Here again, the exact causes are hard to pin down. It is, of course, conceivable that such bias is the result of the advertisers’ own intentions: they would have deliberately chosen to send different ads to men and women. But it is equally as possible that this phenomenon is the result of the algorithm’s reaction to the data it was given. In this instance, as well-known social sciences studies have already pointed out, women are more likely to practise self-censorship. Thus men could have tended to click more often on ads for highly paid jobs. From this point of view, the algorithm’s gender bias would therefore have originated from a pre-existing bias in society.

¹⁹ <https://usbeketrica.com/article/un-algorithme-peut-il-predire-le-risque-de-recidive-des-detenus>
²⁰ <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

A third example: in April 2016, Amazon was found to have excluded from one of its new services (free same-day delivery) predominantly poor neighbourhoods in Boston, Atlanta, Chicago, Dallas, New York City and Washington. At the outset, an Amazon algorithm had, by analysing the data at its disposal, shown that the neighbourhoods in question did not represent profitable areas for the business. Even if Amazon's intention was not to exclude its services from areas because they were predominantly black, this nevertheless ended up being the outcome of this algorithm being put to use. In six large cities, it becomes clear that "the service area excludes predominantly black ZIP codes to varying degrees [...] Black citizens are about half as likely to live in neighborhoods with access to Amazon same-day delivery as white residents"²¹. In Boston, three ZIP codes encompassing the primarily black neighbourhood of Roxbury were excluded from same-day service, while the neighbourhoods that surround it on all sides were eligible.

How can this phenomenon be explained, when Amazon has stressed – quite rightly we don't doubt – that it did not curate any racial data to give to the algorithm? It was put to Amazon that the neighbourhoods in question were exactly the same as those that had, for decades, been subject to a practice known as "redlining". It refers to banks systematically refusing to grant loans to African-Americans – even if they are solvent – solely on the basis of their skin colour and their residence in predominantly minority areas. It is therefore evident that Amazon's algorithm has ended up reproducing pre-existing discriminations, even when there is no intentional racism involved here.

The human configuration of algorithms (i.e. the explicit definition of criteria which determine the way they function and sort, select and recommend) may of course be the source of bias and discrimination. But, as we can see from the three examples above, it is the bias generated by the systems' input data that poses the most daunting challenge today.

By referring to historical patterns, a dataset can reproduce pre-existing discriminations or inequalities. It is therefore

The very people who are curating the input data are unaware of bias, and the users who are its subjects are not necessarily wised up to it either



DID YOU KNOW?

At the debate organised on 24 June 2017 by the Génotoul (which reflects on legal and ethical questions raised by biosciences), Philippe Besse, Professor of Mathematics and Statistics at the University of Toulouse, made the point that we are not all equal when it comes to personalised medicine, as the databases currently used contain considerable bias. One study has shown that, in 2009, 96% of samples taken from these bases had European ancestors (the demonstration bore on 1.5 million samples). Other sources of bias are age (since all such databases tend to be populated by relatively old people) and gender, with several recent publications underscoring the importance of the gender effect on the onset of the diseases in question. The chromosome X is vastly under-represented in these databases, and the Y chromosome hardly features at all. Philippe Besse draws this conclusion: "if you are a young African woman, I think you can consider personalised medicine to be out of reach and pointless".

quite possible that an algorithm seeking to define which profiles should be recruited could exclude women when it is based on a set of profiles corresponding to the most successful career paths within a company in the past. Either because they were excluded in the past, or because they tended to take career breaks more often than their male colleagues for example. Also bear in mind that, for the company in question, irrational use of this kind of algorithm could end up depriving itself of certain talents. The ethical implications would then become directly tangled up with a question of efficiency.

As such, the very operation of training algorithms – through the curation it implies of the data to be taken into account – seems to raise a crucial ethical and legal issue, and not just one of efficiency or of a technical nature. This issue partly overlaps with the one involving the delegation of decision-making, discussed earlier: choosing which input data to use for the training stages clearly entails making decisions that could have far-reaching consequences. But what makes the issue we are talking about here so specific is that it involves making decisions and choices at times in an almost unconscious manner (whereas coding a traditional, deterministic algorithm is always a deliberate operation).

²¹ <https://www.bloomberg.com/graphics/2016-amazon-same-day/>

ration). Whoever trains an algorithm in some ways builds into it his or her own way of seeing the world, values or, at the very least, the values which are more or less directly inherent in the data gathered from the past. Researcher Kate Crawford, in particular, has thus lifted the lid on the ingrained social, racial and gender bias that is rife among the circles where those who are training artificial intelligence today are recruited²².

All of this goes a long way to explaining one of the biggest problems associated with the bias and discrimination that these algorithms can replicate: they are often particularly difficult to detect. The very people who are curating the input data are unaware of them, and the users who are their subjects are not necessarily wised up to them either. The targeted nature of the job ads we mentioned earlier means that the women involved were unaware of the job ads that men were receiving at the same time. This is one of the consequences of the “filter bubble” phenomenon, which we will delve more into later. Finally, artificial intelligence systems make choices, the underlying logic (or even the existence) of which is beyond the grasp of their designers.

All in all, algorithm-generated bias and discrimination raise two key questions today. Firstly, should we postulate that, at least in some cases, artificial intelligence only ever replicates

bias and discrimination that are already ingrained in society? In other words, in this case algorithms would only ever be “vehicles” of bias – repeating without ever actually creating it themselves. In objection to such a standpoint we could, at the very least, argue that the scale at which they are deployed and their potential impact make them ideal tools for addressing discrimination. In other words, with great power comes great responsibility. Not to mention that it is quite possible they could also have a multiplier effect on such bias.

Secondly, **what can we do to ensure we can properly detect this kind of bias which, as we have already explained, can at times pass unnoticed?** Should we distinguish between bias that might be regarded as acceptable and other forms that society simply will not tolerate (such as the forms mentioned above)? And how can we stamp out such bias effectively whilst making sure that algorithms respect the fundamental values that have been democratically established by our societies?

As a final point, we need to highlight a dimension here that we will see crop up again in this report: not just the individual impacts (on a person) but also the collective impacts that algorithms can have. The exclusion of entire neighbourhoods by an Amazon service provides one example in this regard.

Algorithmic profiling: personalisation versus collective benefits

The fact that algorithms are creeping into every area of our lives, especially our online lives, can be explained by the increasing personalisation of content and services. There is a potential downside in this personalisation at the service of individuals. It may indeed undermine some of the inherently collective processes underpinning our society, from the functioning of democracy to the idea of risk-sharing in the economic order. The impact algorithms can have on individuals has been clearly identified and written into the legislation for some time now; but its collective impacts are now raising questions too.

Filter bubbles and loss of cultural pluralism

The implications of the “filter bubble” have been sparking widespread debate ever since Eli Pariser’s publication on the subject²³. They have to do with the idea that the useful tasks performed by algorithms in terms of classifying and filtering the masses of information that we now have at our fingertips would indirectly erode pluralism and cultural diversity. By filtering information, based on the characte-

²² Kate Crawford, “Artificial Intelligence’s White Guy Problem”, *The New York Times*, 25 juin 2016.

²³ Eli Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, New York, Penguin Press, 2011.

ristics of their profiles, it would seem that algorithms are reinforcing individuals' tendencies to embrace only those objects, people, opinions and cultures that conform to their interests, and to reject the unknown.

There are two levels to be considered in relation to the "filter bubble" phenomenon: that of the individual, and that of society as a whole.

At the individual level, the risk is that each person sees him or herself compared, purely and simply, associated with a digital alter ego. One that is implied from his or her personal data, isolating him or her within a bubble of recommendations that always conforms to this profile. The promises of an access to a greater amount of cultural content than ever before would thus end up being paradoxically cancelled out by individuals' poor exposure to cultural diversity in practice. This could occur even when the individual is looking for such a diversity in principle. French Ministry of Culture (DGMIC) highlights, in this connection, that "algorithm-driven recommendations are based on users' actual consumer practices rather than their desires or aspirations".

It is nevertheless important to point out that many specialists, researchers and practitioners in the digital sector either contest the idea of filter bubbles or are calling for questions to be asked in a more specific way in their regard. Accordingly, Antoinette Rouvroy says: "this matter of the filter bubble is not unique to algorithms: we are highly predictable beings, who behave in a very regular way, and this makes it that much easier to isolate each of us inside bubbles. But we only isolate ourselves if there is profit to be gained. It's all a question of algorithm configuration. They can also, on the contrary, expose us to elements or information that we would never have searched for" (statement from the public debate launch on 23 January 2017 at the CNIL). Admittedly, there is no evidence of this alternative being greatly exploited in practice. Cultural consumer trends are based on a two-sided preference structure: on the one side, strong links "reflecting a proven preference for a previously well identified type of content"; and on the other, weak links "conveying a potential affinity for a type of content that awaits discovery"²⁴. Now, most of the predictive algorithms used by the prominent platforms providing entertaining and cultural services (Netflix, Amazon, Spotify, etc.) focus on the strong links. None of the major categories of algorithms consider serendipity to be a key variable of consumer choices.

It would seem that algorithms are reinforcing individuals' tendencies to embrace only those objects, people, opinions and cultures that conform to their interests, and to reject the unknown

Dominique Cardon, meanwhile, asserts that "digital technology has brought with it a diversity of information without precedent in the whole history of humanity. It is absurd to say that Facebook isolates people. But there are risks involved: curious people will give off curiosity signals and be encouraged in their curiosity in return. On the other hand, people who show little sign of curiosity will be steered towards less diversity. [...] There is a risk that, within a certain context and for a certain audience, social practices develop where the algorithm is not a factor of enrichment or discovery, but rather gives more of the same" (statement from the public debate launch on 23 January 2017 at the CNIL). Lastly, the French Ministry of Culture (DGMIC) points out that incentive to gain a competitive edge and "a liberal-individualistic perspective [from which] diverse exposure can be valued simply because it extends individual choice and affords individuals more opportunities to realize their interests"²⁵ could limit the threats to diversity by urging stakeholders to take up the challenge of filter bubbles and find solutions.

At society level, taken as a whole, the different ways in which individuals are shielded from otherness, from opinions that differ from their own – in terms of politics in particular – could be problematic for the quality and vitality of public debate; for the quality and diversity of information; overall to the healthy functioning of democracies.

Personalisation of information could lead to an extreme fragmentation of the public space and the disappearance of a minimum core set of information shared by the whole of the body politic (which enables the organisation of a proper debate).

²⁴ Report by CSA Lab

²⁵ Natali HELBERGER, Kari KARPPINEN & Lucia D'ACUNTO, "Exposure diversity as a design principle for recommender systems", Information, Communication & Society, 2016.

In an age where increasing numbers of citizens are using social media as their main (and sometimes, the only) means of getting information²⁶, the stakes are high for the future of democratic life. It is true that we tend to surround ourselves with people of similar mind and value. But at least the traditional press, with its editorial policy, informs the reader more clearly on the leanings of the content he is reading. Current debates on the subject nevertheless underline the fact that “filter bubble” effects are not inevitably and, in all cases, produced by algorithms. They are above all the result of how algorithms are configured: they could just as well be programmed differently and be given on the contrary the objective of exposing individuals to strong cultural, informational and political diversity.

The very nature of the problem may well have delayed its arrival in the spotlight. It is indeed quite possible for an individual to live inside his or her bubble of information without even being aware of its existence. Filter bubbles are not likely to be challenged since individuals feel way more comfortable in the absence of contradiction, partly due to the confirmation bias characterising the human mind (with which the cognitive sciences are well familiar). In other words, nothing predisposes an individual to notice that he is caught inside such a bubble. It therefore comes as no great surprise that critical statements regarding “filter bubbles” are often accompanied with stories of the moment of awakening, experienced as something of a shock. The debates on the filter bubble and its political outcomes had substantial media coverage during the 2016 US presidential elections and the Brexit referendum result a few months earlier. Two electoral earthquakes in which more than a few Internet users who supported Hillary Clinton or opponents to Brexit were particularly shocked to find results to which their newsfeed had given no clue. More recently, in August 2017, sociologist Zeynep Tufekci, who specialises in online social movements, was not alone in noticing that her Facebook newsfeed made no mention of the story surrounding Ferguson, even as she could see the hashtag Ferguson spreading on Twitter.

Individuals' hazy grasp of the underlying workings of the platforms they use to get information might be considered a big part of the problem. One study has thus shown that more than 60% of Facebook users are completely unaware of the editorial activity that the algorithm actually carries out, and think instead that every single of their friends' posts and pages they follow appear in their newsfeed.²⁷

In reality, they are only shown 20%, selected according to several factors: advertising of the post, past interaction of the user with posts that are considered similar – *like*, comment, share –, number of other users having done the same, and so on.

The use that the digital economy makes of algorithms for the purposes of personalising services and user experience therefore follows a mindset that poses a problem when its effects are considered from no longer just an economic point of view, but also a cultural or political one. **The end goal of the main platforms using algorithms is consumer satisfaction; the consumer being understood as a *homo economicus*. The large-scale political and cultural effects of their algorithms are only secondary details.**

Atomisation of the political community

This effect brought about by algorithms and their personalisation function can nevertheless become a direct lever for certain stakeholders intent on exploiting them to wield influence or even to manipulate. Fake news, which hit the headlines during Donald Trump's campaign – although not directly resulting from algorithms – are spread and gain traction within echo chambers formed by social media or search engine algorithms. More directly still, very sophisticated political campaigning software is now able to target voters more precisely than ever. This leads to a potentially unprecedented fragmentation of political messages, henceforth crafted for the attention of atomised individuals. The practices of Cambridge Analytica, the firm which did digital work for Trump's presidential campaign, are at the very forefront of these new uses being made of algorithms for electoral purposes (see inset). The increasing tailoring of the political narrative to align with individuals' profiles, thanks to AI's growing capacity to curate messages based on these profiles, is prompting serious questions. Should we see this as a form of manipulation? Should we be laying down limits in this regard? Should we view these practices as inevitable side effects of the technological shift which, because they are difficult to regulate, require us to think up ways to counterbalance them? If so, how can we go about this?

As we can see, the theme of the filter bubble is the flip side of algorithmic personalisation. This means that filter bubbles and fragmentation can also occur in other sectors than cultural consumption and media or politics.

²⁶ According to the Pew Research Center, 61% of Millennials “report getting political news on Facebook” (Pew Research Center, *Millennials & Political News. Social Media – the Local TV for the Next Generation?*, June 2015).

²⁷ http://www-personal.umich.edu/~csandvig/research/Eslami_Algorithms_CHI15.pdf



FOCUS

Algorithms and political campaigning

Election campaign software, based on the use of predictive algorithms for analysing electoral data, was increasingly harnessed during the most recent presidential elections, in the United States as well as France. A far cry from the more traditional campaign methods, highly targeted political messages can now be conveyed to voters. The most accomplished examples of such individual profiling can be identified in the United States. Already during the 2008 and 2012 presidential elections, Barack Obama's campaign teams had scores of datasets at their disposal on virtually all voters. In 2016, by analysing data harvested from social media sites and data brokers, Cambridge Analytica could have sent out thousands of extremely personalised pro-Trump messages in the space of just one evening²⁸. Although this company has subsequently struck a tone aimed at downplaying its initial claims, this incident still lifts the lid on an underlying trend that is likely to grow in the future.

In France, the **principles governing personal data protection** nevertheless limit the extent to which such individual targeting software can develop in practice, since consent is a prerequisite for such data collection. Incidentally, in a press release from November 2016, the CNIL gave a reminder of the rules for using social media data for political communication purposes²⁹.

Filter bubbles: a cross-cutting challenge

The questions raised by filter bubbles are not exclusive to the culture, information and political sectors. Prediction and recommendation functions at work in the algorithmic systems, embedded in the digital ecosystem today, are likely to spawn self-fulfilling prophecies that can enclose data subjects within a "predicted" destiny.

Is a form of isolation not a possible consequence of future uses of learning analytics and adaptive learning? Without casting doubt over the promises of such techniques, it is quite reasonable for us to question the possible effects of systems claiming to define learning paths on the basis of each student's profile, and of predictions established through the application of a mathematical model to this profile. Is there not a risk that the prediction becomes self-fulfilling and that the student sees his or her academic and professional future mapped out for them, as soon as the analysis has been made?

As highlighted by Roger-François Gauthier, "with learning analytics, prediction could lead to pupils being confined within certain pathways. In France, not enough attention is being paid to this problem – when we need to make sure students are not subject to determinism, and for that, the question of the values written into algorithmic systems is fundamental³⁰".

In the same way, it is possible to associate with the filter bubble idea some potential impacts of algorithm use in the human resources and recruitment sector. Laurence Devillers thus speaks of the risk of "standardisation of profiles" that an algorithm – or imprudent use of the algorithm – might pose for the recruiter. In some ways it would be the latter that would fall victim to a bubble containing solely those profiles predefined in advance. He would be deprived of the measure of serendipity inherent in the recruitment process insofar as this can bring to the fore profiles which, whilst not a match to the criteria set beforehand, end up having much to offer. How can such profiles be spotted if a growing part of the selection of candidates is being delegated to automated systems?

²⁸ <https://www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage>

²⁹ <https://www.cnil.fr/fr/communication-politique-queles-sont-les-regles-pour-lutilisation-des-donnees-issues-des-reseaux>

³⁰ Statement from the public debate launch on 23 January 2017, at the CNIL.

Demutualisation

Algorithmic personalisation poses a specific challenge to the insurance sector. **This trend towards ever more personalised services would seem to be calling into question the very principle of mutualisation on which insurance and its social pact are founded.** That a group of individuals agree to insure themselves, i.e. share their risks, supposes that these risks remain at least partially unknown to them. I insure myself without knowing who, out of me or my neighbour, will contract a disease incurring significant health expenses. But the greater segmentation enabled by the use of the masses of data generated by individuals' behaviours online (social media in particular) or off line (data harvested from smart wristbands for example) would have a tendency to lift the "veil of ignorance"³¹ underlying the pooling of insurance risks and which a basic level of segmentation helps to maintain.

Might such innovations not beget new forms of discrimination and exclusion? People who are deemed to be "at risk" could be lumbered with higher rates, or even be denied insurance cover altogether. Furthermore, associating a type of behaviour with the risk of developing a particular illness could end up penalising individuals adopting a lifestyle that is deemed to be "at risk" (such as smoking or following a diet that is considered to be too high in fat or sugar).

The question would then bear on how to limit what can come across as an excessive standardisation of individual behaviours when these would be considered "unhealthy". The risk would be that algorithms, via the correlations they make in datasets, end up laying down the one and only accepted set of norms for individual behaviours (from which we could only deviate by paying a higher insurance premium). Unlike a choice such as raising the price of tobacco for example (smoking is considered to be a cost for the community), such decisions would not involve any collective deliberation, but result directly from the data input. What is more, an approach of this sort would completely do away with the collective and social determinants of behaviours, by focusing solely on the accountability of individuals. Other risk factors, associated with the individual's environment or genetic makeup, would likely lead to inevitable discrimination and exclusion insofar as these are completely out of the hands of the individuals in question.

Although a race for the "right risks" could therefore escalate between insurance companies, it is doubtful that this is a good thing for the latter as a whole. There are merits to mutualisation for insurers. According to Florence Picard, of the Institut des Actuaire (French Institute of Actuaries), *"the more strictly groups are segmented, the greater the risk to mutualisation. The aim is that the risk can be controlled: the greater the segmentation, the greater the room for error"*³².

Preventing massive files while enhancing AI: seeking a new balance

The algorithms we use on a daily basis work by processing scores of data, including a significant proportion of personal data: digital traces left by our online browsing, by using our smartphones, our credit cards and so on. **The quest for ever better performing algorithms is calling for the increasing collection, processing and retention of personal data.**

We might therefore wonder whether the development of artificial intelligence might, sooner or later, run counter to the ethical principles enshrined in the law since the 1978 French Data Protection Act. Artificial intelligence consumes reams upon reams of data; it requires a large memory (i.e. databases which it will retain for as long as possible). The principles of the 1978 Act refer, by means of the principle of specifying a purpose, to a minimisation of personal data

³¹ Here, Antoinette Rouvroy applies the concept, proposed by John Rawls to establish a thought experiment for considering a moral problem, to the insurance sector.

³² "Algorithmes et risques de discriminations dans le secteur de l'assurance" (Algorithms and risks of discrimination in the insurance sector), event organised by the Ligue des Droits de l'Homme (Human Rights' League) on 15 September 2017.

collection as well as the limitation of the retention period of said data as necessary safeguards for the protection of data subjects and their freedoms.

The principles of the 1978 Act (also set forth in the General Data Protection Regulation which is poised to come into force in May 2018) provide a general balance which allows a certain amount of flexibility overall. One example: heightened security measures can, to a certain extent, be considered to offset a longer data retention period. That said, there is the possibility that the sheer scale of technological change brought about by the development of artificial intelligence calls this state of affairs into question.

For example, the leaps and bounds being made in precision medicine seem conditional upon the compilation of ever larger databases, both in terms of the numbers of data subjects and of the amount and variety of data retained on each of the latter. In this way, epigenetics thus claims to combine an approach relying on individuals' genetic data with an approach that takes environmental data into account – i.e. data concerning the setting and even the lifestyle of the "patient" (assuming that this notion still means something in a context where the focus is increasingly on "prediction"). The promise of predictive medicine is, by obtaining as detailed a profile as possible of an individual and of his or her disorder, to be able to compare it

to other individuals with similar profiles. The objective: identifying the most appropriate treatment for this patient. We could go so far as to maintain that the healthcare benefits pursued entails the compilation of vast databases. Yet there are no guidelines stating where this data collection should stop: the medical record? The genome? Epigenetic data – i.e. environmental data (on living habits, living environment, diet, etc.)? Going back how many years? Note that this type of dilemma is in no way exclusive to medicine. It is also an issue to be addressed from a similar angle in security policies for example, where the requirement to identify suspects seems to justify the collection of bigger and bigger data on individuals.

It is clear that **the question here concerns the balance to be struck between protection of freedoms (protection of personal data) and medical progress**. This report does not seek to determine what this balance should be, inasmuch as this warrants an in-depth discussion – one that would have to include an assessment of precisely how much progress is expected to be made in precision medicine. Philippe Besse, Professor of mathematics at the University of Toulouse, believes that the data made available to medical research under the National Health Data System (SNDS) is sufficient for making progress which will, in any case, be limited by the complexity of the living organism – to a level far below the hype of certain prophecies³³.

Quality, quantity, relevance: the challenges of data curated for AI

Algorithmic systems and artificial intelligence are dependent on the input data (whether personal or otherwise) they are given and which they process to produce an output. In short, this characteristic raises three associated, albeit separate, challenges: bearing on the quality, quantity and relevance of the data supplied to these systems.

The matter of the quality of the data processed by algorithms and AI is the most straightforward. It is not difficult to understand that **incorrect data or data that is quite simply out of date will lead to errors or malfunctions of varying gravity depending on the sector in question**, from the mere dispatch of targeted advertising that does not

match my actual profile, to an incorrect medical diagnosis. Ensuring the quality of input data in algorithmic and AI systems is thus a challenge that is set to take on ever greater importance as these machines become ever more autonomous. But this is a costly challenge. Data corruption might just as well result from a very tangible technical glitch caused by damaged sensors collecting this data, as from a human problem stemming from the interest some stakeholders introducing bias in the input data. The temptation for negligence in this regard must be taken seriously. Especially in some areas where the impact of poor quality data might not be immediately perceptible, such as human resources and recruitment. In this regard, the reliability of

the data available on professional social networks must not be regarded as an inexhaustible resource, and must instead be questioned (given people's tendencies to embroider their CV or, alternatively, to the lack of updates). Such negligence also comes from the confidence that the user has in the output produced by a machine, deemed to be objective and more competent than humans.

The quantity of available data can be another factor detrimental to the quality of the output produced by algorithms and AI systems. Cathy O'Neil refers in this regard to the example of a local authority in the United States which used an algorithm-based program to assess teachers. Use of this program resulted in the laying-off of teachers who turned out to have a good reputation within the local communities where they worked. One of the main reasons is that the algorithm used to assess the annual progress of each teacher would need much more than the data concerning a few dozen students at the most. Such a limited number of cases cannot have any statistical value in a situation where there are many different variables likely to explain, on the one hand, the teacher's performance and, on the other, the poor grades of a student (relationship problems, family problems, health problems, etc.). The only value to be found in this result is that it gives the decision-makers the sense that they are making rational, objective and effective decisions, the prestige of the machine being a pretext.

This does not in any way mean that the collection of data should constitute an end per se. In some cases there would be more value in having a variety of data than simply a large quantity. In the case of the algorithm of a GPS application for example, the data pertaining to millions of vehicles following the same route will be of less use than a much smaller set of data from vehicles travelling along a greater variety of routes.

Lastly, **the question of data relevance has less to do with the truthfulness of this data and more with the bias that can be introduced when it is curated.** As has already been shown (See "Bias, discrimination and exclusion"), it may be entirely true that very few women lead an uninterrupted high-level career in a particular company. But taking this fact as a sign of women's ability in the future to accomplish successful careers in this same company plainly amounts to a biased approach. In this case, the dataset in question incorporates forms of inequality and/or discrimination. Ignoring this type of bias would equate to allow such phenomena to keep going.

What becomes apparent through these three challenges is that the promise held by algorithms can only come to bear if the utmost stringency is practised in the collection and processing of the data used. That such stringency (along with investment in material and human resources) may be overlooked by certain stakeholders poses an evident risk, when algorithms are often described as being sources of "objective" or "neutral" truth. In the example of the algorithm used to assess teachers in the United States highlighted by Cathy O'Neil, **the methodological negligence on the part of the algorithm's designers and promoters has resulted in the users placing excessive trust, devoid of any critical thinking, in the algorithm** (by focusing solely on the need to obtain a quota of teachers to be dismissed from the system). Yet, although ensuring the quality and relevance of the input data given to algorithms thus strikes as an ethical requirement, ultimately this is also a condition for the sustainable utility of algorithms for users and for the wider society.



SURVEY

People become more sceptical of algorithms and AI with age *

Young people have more faith in the opportunities harboured by algorithms: 68.5% of 18-24 year olds believe that the opportunities outweigh the potential threats. However, only 36% of 55-64 year olds consider the benefits to be greater than the risks.

Some algorithm applications are viewed more favourably by youngsters: 75% of 18-24 year olds think recommendations are a good thing for online purchases (versus 48% for the whole panel), and 50% when choosing a soulmate (versus 26%).

* Survey carried out as part of the public debate by the rural-based family association "Familles rurales", among 1,076 of its members.

Human identity before the challenge of artificial intelligence

The idea of an irreducible human uniqueness is being challenged by the autonomisation of machines on the one hand, and the increasing hybridisation of humans with machines on the other.

Ethical machines?

The boundaries between humans and machines firstly begin being questioned with the concept of "ethical machine". It appears as a radical way to address the questions raised by the potential delegation of decisions to autonomous machines (artificial intelligence). Making machines "ethical" would be one solution to the problems touched on earlier in this report. Such a line of thinking probes the question of whether it is even possible to establish an ethical framework³⁴ for programming into a machine. In other words, **is it possible to automate ethics?** This emerged during the debates as one of the key issues currently grabbing the attention of the community of researchers in artificial intelligence, as highlighted by Gilles Dowek (CERNA) during the study day organised at the "Collège des Bernardins" on 20 September 2017.

The famous trolley problem frequently crops up during discussions on this subject. We know that this dilemma involves a trolley with no brakes hurtling down a slope; the trolley arrives at a junction; depending on which of the two tracks it goes down, it will either kill one, or several people. How should a person with the possibility of pulling a lever to switch tracks and therefore choose, as it were, one of the two possible scenarios, react? The interest of this thought experiment is that it can give rise to a whole host of variants: what if the lone person bound to one of the two tracks turned out to be a close relative? Or if there were 5, or 100, people on the other track?

It is easy to see how similar dilemmas can be applied to autonomous vehicles that are soon expected to be taking to our roads. According to what principles should a car faced with an ethical dilemma of this type "choose" to react? The trolley problem reveals to what extent different "ethical" choices are possible. Where this kind of situation would

have been anticipated at the stage of the system's development, it would of course be possible to provide them with an answer. But, surely **what is specific about ethics is precisely that it concerns previously unencountered situations, possibly involving conflicts of values and the solution to which must be worked out by the subject** (the play *Antigone* comes to mind, where the ethical conflict is between family loyalty and civic duty). Isn't the point that it is always worked out in the heat of the moment? And doesn't this therefore make the whole hypothesis of establishing a pre-defined ethical framework somewhat delusive? At the very least, it implies an implicit conception on our part which is far from straightforward.

Let us just say, for the time being, that expressions such as "ethics of algorithms" or "ethical algorithms" should not be taken literally. They contain a measure of anthropomorphism; since they attribute human capacities to machines. Some are of the opinion that these expressions risk skewing the debate; which should instead be focusing on the requirements to be met by humans (those who design, train, roll out and use algorithmic systems and artificial intelligence).

They should therefore be seen merely as a handy metaphor that is not to be taken literally. However, as pointed out by Gilles Dowek for example, using this type of metaphor can be considered entirely justified insofar as it acknowledges the growing autonomy of these systems and the need to establish, to the extent possible, an ethical framework to be programmed into algorithms. At the end of the day, even if it were possible to code ethics as is into a machine (meaning if this machine were able not only to respond in a certain way to an ethical situation envisaged beforehand, during the development stage, but also to tackle new situations by applying ethical thinking to them), the choice of which ethics to be coded would remain the responsibility of humans. The real challenge, then, is to make sure that the ethical choices made at the development stage are not commandeered by "a small caste of scribes" (Antoine Garapon). The sheer scale on which algorithms are being rolled out in this digital age makes this a democratic question of the utmost importance.

³⁴ Meaning a general rule for assessing how to respond to any situation – deontology or consequentialism – or a set of rules fulfilling the same role – Kantian ethics or Buddhist ethics for example.

The hybridisation of humans and machines: rethinking human identity?

One way to consider the ethical question applied to algorithms and artificial intelligence might be to view these in light of the statement – set out in Article 1 of the French Data Protection Act – that information technology “shall [not] infringe human identity”.

This report has previously looked at problems associated with the way humans organise their action with machines – an age-old question that has gained fresh currency with the emergence of ever more “autonomous” machines in this age of algorithms and artificial intelligence³⁵. This point underscores the fact that, depending on how these technologies are developed, one of the components of human identity and dignity – namely our freedom and responsibility – may well be impacted. The rise of a form of machine “autonomy” obviously needs to be carefully qualified. Gérard Berry, Professor at the “Collège de France” and “Algorithms, machines and languages” chairholder, puts it this way: “one day, we are told, machines will talk and be autonomous, and digital technology will give rise to a new form of life. No one is saying when machines will gain their autonomy and capacity for creative thinking, and I don’t know this either – not by a long way. Above all, what kind of life are we talking about?³⁶”. But for all that, we could ask ourselves whether the technological course on which we are already set should not be prompting questions over the relevance of the notion of “human identity” itself (insofar as this implies a watertight separation between human and non-human). With the issue of a “legal status for robots” already raised by legal experts and recently examined by the European Parliament (Delvaux report) comes the prospect of this possible blurring of the lines of what constitutes human. In response to such post-humanist arguments, humanist tradition could admittedly fire back that machine autonomy is nothing more than an illusion today. It is solely a metaphor intended to des-

Depending on how these technologies are developed, one of the components of human identity and dignity – namely our freedom and responsibility – may well be impacted

cribe a complex object which ultimately masks very real human liability and action – however watered down and fragmented these may have become.

The beginning of an hybridisation between humans and machines is taking place in terms of action, but our attention is also called to broaden in the future to factor in the upcoming physical form of hybridisation between algorithms, humans and even animals (through smart and communicating implants). This physical hybridisation is another stage in the evolutionary path along which we are already bound in the ongoing interaction now linking us to a whole host of algorithmic processes.

Finally, this theme of an unclear boundary between humans and things (or rather, between humans and machines) has already come to crystal-clear fruition in the phenomenological context of certain recent robotic application trials which intend to give robots the appearance of humans. One example is the robot Pepper by the firm Aldebaran, designed to be used in shopping centres to interact with customers. Above all, and this is directly tied in with the subject of algorithms and artificial intelligence, **an entire research field is geared towards designing empathetic robots capable of perceiving human emotions** (by analysing the face or the voice for example) so as to adapt their behaviour to the mood of their interlocutor. Such research raises the question of where the limit lies between, on the one hand, the benefits of a form of AI capable of understanding and adapting to the moods of its interlocutors and, on the other, a form of manipulation relying on technical engineering that is capable of exploiting our emotional vulnerabilities³⁷. A second question, related to the first, is to know to what extent the capacity for illusion unique to these technologies, and the imbalance that will exist between these robots and the people whose emotions they will read, make them morally acceptable? Sherry Turkle, Professor at MIT, maintains that humans readily attribute a subjective and sensitive side to robots³⁸. And there is a strong temptation for aging societies to increasingly entrust care of the elderly to this type of robot. In France, Serge Tisseron has been devoting critical thought to these technologies³⁹. Whatever the answers found to these questions, it seems essential that they in no way hide the societal choice and political dimension to be weighed up in the decision to use robots to assist the vulnerable members of our societies, instead of investing in other types of resources (time, staff and so on).

³⁵ The question of hybridisation between humans and artefacts is not new (Socrates was already commenting on it in Plato’s *Phaedrus*): algorithms are helping to shape our identity in the same way that writing affects our memory skills and constitutes a silent artefact, incapable of the slightest explanation. That the idea of a strictly separate “human identity” from objects is being questioned does not, therefore, necessarily imply a radical new concept.

³⁶ Gérard Berry, “Non, l’intelligence artificielle ne menace pas l’humanité!”, interview in *Le Point*, 18 May 2015.

³⁷ There are strong parallels between this issue and the one raised by political communication models which are supposed to tailor the candidate’s message to the expectations of each targeted and profiled individual.

³⁸ Sherry Turkle, *Alone together*, New York, Basic Books, January 2011.

³⁹ Serge Tisseron, *Le Jour où mon robot m’aimera. Vers l’empathie artificielle*, Paris, 2015.

How can we respond?

From ethical thinking to algorithmic regulation

P.44

What the law already says about algorithms and artificial intelligence

P.45

The limits of the current legal framework

P.46

Should algorithms and artificial intelligence be banned in certain sectors?

P.47

**Two founding principles for the development of algorithms and artificial intelligence:
fairness and continued attention and vigilance**

P.48

Engineering principles: intelligibility, accountability, human intervention

P.51

From principles to policy recommendations

P.53

How can we respond?

From ethical thinking to algorithmic regulation

Should algorithms be regulated?

This question has cropped up several times over recent months in both the general print media and among experts in the fields of digital technology and public policy.

It is in fact merely an extension of the question of digital regulation itself. We know that the digital environment has partly arisen in opposition to the idea of standards – or legal standards at any rate. Evidence is given by the counterculture that developed in the United States in the 1960s or digital firms' insistence that innovation must not be hindered by a system of standards unsuitable to reality. This distrust in regulation is a common thread running through the past few decades. One of the clearest formulations of this view can be found in John Perry Barlow's famous Declaration of the Independence of Cyberspace from 1996. But for a number of years now, such thinking has had to contend with state actors' efforts to subject the digital environment to ordinary law, sometimes mechanically and at other times by deploying full-on legal innovations.

Many stakeholders today maintain that algorithms and artificial intelligence should not be regulated. They argue that it would be too soon to enforce rules that are bound to prove unsuitable. They would quickly be obsolete because of the breakneck speed at which technological progress is taking place. Legal invention could simply in any case keep pace.

Truth be told though, such a position ignores a legal reality that is as far-reaching as it is at times little known: **algorithms and their uses are already governed, whether directly or indirectly, by a raft of legal regulations.** Admittedly, as we will see, these rules are scattered across diverse laws and codes in practice, reflecting the cross-cutting nature of digital technology.

Polls carried out during the public debate initiated by the CNIL incidentally revealed an expectation for rules and limits when it comes to algorithms and artificial intelligence. Such rules and limits may take other forms than just binding standards, for example soft regulation such as "charters" adopted by a company, by a profession or by a sector. This came across in the survey conducted by the CFE-CGC, among 1,263 of its members, for example⁴⁰.

The setup by Parliament of a discussion assignment entrusted to the CNIL on the ethical and societal issues raised by the development of digital technologies fits squarely into this context. It is a clear sign of a commitment to think about the limits, about the standards – whatever form these might come in – to be set for new technology. And it also shows that the public authority does not intend to give into the temptation of rushing the regulation process and ending up with rules that do not meet requirements. In this regard, the belief that, alongside the emergence and uptake of new technology, thought must be given to its limits does not in any way mean that the law systematically represents the right approach to laying down these limits. This was the CNIL's perspective in any case, hence why it wished to open up the debate as widely as possible, not only to include public stakeholders, but also practitioners, professionals and the general public.

To be able to draw up recommendations, we therefore had to begin by exploring the main innovations, and the ethical and societal issues that these raise. This has been the subject of the first parts of our report. The pages that follow will set out to review the main principles that are likely to address these issues as well as the concrete recommendations we are in a position to make today.

⁴⁰ To the question "Do you consider it a priority to define an ethical charter on the use of algorithms in HR management and recruitment?", 92% answered yes.

What the law already says about algorithms and artificial intelligence

Not all of the challenges identified in this report are new.

The Tricot Commission, whose report formed the basis for the French Data Protection Act of 1978, had already drawn attention to some of them. The discussion, over and above data processing, bore on the challenges raised by the computerisation of the State and French society. The risk of discrimination or exclusion of people, as well as the risk of excessive trust being placed in computers, have been clearly enunciated from the beginning, along with other challenges directly connected with the ability to collect and store huge amounts of data. The debate over whether or not there is a need to “regulate algorithms” quite simply overlooks the fact that algorithms have been governed by legislation (the Data Protection Act in particular, but other laws too) for some forty years already.

The debate over whether or not there is a need to “regulate algorithms” overlooks the fact that algorithms have been governed by legislation for some forty years already

These three principles are also laid down in the General Data Protection Regulation (GDPR) due to come into force in May 2018. They are as follows:

First, the law governs the use of personal data required for the operation of algorithms, beyond the strict stage of algorithmic processing. In other words, it governs the conditions for collecting and retaining data⁴¹, as well as the exercise of data subjects' rights (right of information, right to object, right of access, right to rectification) in order to protect their privacy and freedoms.

Second, the Data Protection Act prohibits a machine from being able to make decisions alone (with no human intervention) where there are significant consequences involved for the data subjects (court judgment or decision to grant a loan for example)⁴².

Third, the law provides that data subjects shall have the right to obtain from the controller information about the logic involved in algorithm-based processing⁴³.

Beyond the Data Protection Act, other, older pieces of legislation provide a framework and series of limits for the use of algorithms in certain sectors, precisely insofar as they regulate these sectors⁴⁴. The question of algorithmic collusion taxing competition regulators today, for instance, is not completely devoid of legal references: it has more to do with the effectiveness of the rule of law and the need to invent new ways of proving the existence of unlawful conspiracy⁴⁵.

Indeed, the 1978 Data Protection Act, in which the Tricot Commission's work culminated, contains a certain number of provisions that could be summarised according to three principles – themselves coming under a single general principle enshrined in Article 1: “data processing shall be at the service of every citizen. It shall develop in the context of international cooperation. It shall infringe neither human identity, nor the rights of man, nor privacy, nor individual or public liberties”.

The legal provisions prohibiting different forms of discrimination, drawn up in the wake of Article 7 of the Universal Declaration of the Rights of Man, can also be applied naturally to algorithms⁴⁶.

⁴¹ Principles of purpose, proportionality, security and limitation of the data storage period.

⁴² Article 10 of the 1978 Act, Article 22 of the GDPR.

⁴³ Article 39 of the 1978 Act. Article 15.1 (h) of the General Data Protection Regulation (GDPR) provides that the data subject shall have the right to obtain from the controller the following information: “the existence of automated decision making including profiling referred to in Article 20(1) and (3) and at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject”. The legal limits laid down by the GDPR particularly concern “profiling” (no decision based solely on processing, subject to certain exceptions).

⁴⁴ With a little effort, we could consider how the Public Health Code (which punishes the unlawful practice of medicine by anyone without a medical qualification) might apply to artificial intelligence-based systems in the medical sector. We could also imagine it being unlawful for an algorithm to make a diagnosis alone, pursuant to this legal provision. This legislation came about at the turn of the 19th century as a response to the authorities' determination to crack down on “charlatanism”. Critics of the exaggerated promises made by some companies will not miss the ironic parallels with the current situation.

⁴⁵ <http://internetactu.blog.lemonde.fr/2017/02/11/comment-prouver-les-pratiques-anticoncurrentielles-a-lheure-de-leur-optimisation-algorithmique/>

⁴⁶ “All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination”.

The limits of the current legal framework

A certain number of issues raised by algorithms nevertheless represent a blind spot of law and the various aforementioned legal provisions to date.

Focus on algorithms that process personal data and no consideration of the collective effects of algorithms

A first point to make is that **these provisions concern algorithms only to the extent that they use personal data for their tasks** and that their output applies directly to data subjects. This is particularly the case of the Data Protection Act, the only one among the French pieces of legislation we mentioned to bear directly on algorithms (referred to as “automated processing of personal data”). Yet many algorithms do not use personal data. Trading algorithms are one example. The impacts of these algorithms that process non-personal data are just as likely to raise questions as the others are. Trading algorithms concern a sector which is highly regulated incidentally, but other examples yield an insight into the possible impacts of algorithms that do not process personal data. The example cited earlier on in this report (see “A question of scale: the massive delegation of non-critical decisions”) of Cathy O’Neil’s algorithm dreamt up to plan her children’s meals enables her to shed light on the specific challenges associated with the scale of the impact of algorithms run by computer systems. We could also imagine a nationwide algorithm aimed at planning the menus in school canteens on the basis of certain criteria (cheaper products or nutritional quality for example). Although it does not process any personal data, this kind of algorithm could have major social and economic implications simply because of the scale on which it is deployed. But the law provides no guidance on this new dimension to date.

Second, **the legal provisions referred to above concern the effects that algorithms have on data subjects – i.e. from an individual perspective – but make no direct mention of any collective effects.**

Think of the impact that algorithms used for electoral marketing can have on democracy itself (See: “Atomisation of the political community”). So although the Data Protection Act might be regarded to constitute a limiting factor in terms of such impacts⁴⁷, this is only indirectly speaking, without being its primary objective.

The limits to the law’s effectiveness

Another type of limit in the regulation of algorithms and AI that can be identified in current legal provisions has to do with the very effectiveness of these and the principles they are supposed to uphold. In a fast-moving digital world, where there is a strong imbalance between those who supervise algorithms and data, and the data subjects, the latter struggle to exercise their rights (such as the right to obtain human intervention in the context of a decision made based on algorithmic processing, or the right to obtain information about the logic underpinning the operation of the algorithm).

A series of recent debates have been held with a view to taking this reality into account, and some have given rise to new pieces of legislation. The General Data Protection Regulation (applicable from May 2018) provides several answers to this question on the effectiveness of law in the digital environment – including where algorithms are concerned⁴⁸. The Digital Republic Bill was adopted in October 2016 with this same ambition to strengthen the effectiveness of pre-existing principles in mind.

On the one hand, it has strengthened controllers’ obligation to inform data subjects when algorithms are at play. On the other, it stipulates that the source codes of algorithms used by government departments must be communicable documents. This thereby more firmly establishes (with the notable exception of the private sector) the right to obtain information on the logic involved in algorithmic processing, enshrined in the 1978 Act.

⁴⁷ The Data Protection Act particularly makes the enrichment of individual profiles by data harvested on social media conditional upon obtaining the data subjects’ consent.

⁴⁸ Article 14.1a of the GDPR, for example, strengthens the right of information by providing for clear and intelligible information supplied by the algorithmic processing controller of his own accord.

Should algorithms and artificial intelligence be banned in certain sectors?

The question of whether or not algorithms and artificial intelligence should be banned in certain sectors or for certain uses should not be left out of a discussion on the ethical issues these technologies raise. At the launch event organised for the debate by the CNIL on 23 January 2017, Rand Hindi thus asked whether we should draw the line at automating certain occupations on ethical grounds.

The highly sensitive nature of a certain number of sectors and the decisions that are made within them means that, fairly logically, these are the areas where the question of such bans could be raised. Accordingly, in the military sector there has been an international petition recently calling for autonomous weapons to be banned. Medicine and justice are other sectors where this question might be asked. Granted, as has already been pointed out, the legislation already stipulates that automated decision-making is not allowed when it comes to a doctor's diagnosis or judge's decision. Given the still somewhat blurred line between support for decision-making and delegation of decision-making, the question of a solemn reminder of these principles seems entirely relevant.

There are also calls for a ban in other sectors which do not initially strike as so sensitive. Serge Tisseron has recently taken a stance against personalised targeting in the advertising and culture fields for example, which he accuses of "dooming each spectator to going round and round in terms of what they know of their likes and what they don't know of their preconceptions". Such targeting would partly be responsible for "reducing the volume of data that most humans have at their disposal to form an opinion about the world⁴⁹".

On a final note, the ban applied to a particular use of algorithms could bear on the data used, similar to the moratorium declared by French insurance companies in 1994 on the use of genetic data, and renewed in 2002 by the Kouchner Act. Again in this sector, limiting the use of data might also be a solution (through the Law or through initiatives by the stakeholders themselves) for maintaining the essential "veil of ignorance" to continuing the system of sharing risk.



WHAT THE PUBLIC THINK

Participants in the public consultation which the CNIL organised in Montpellier on 14 October 2017 (see "Organisation of the public debate on the ethical issues of algorithms and artificial intelligence") identified a certain number of ethical issues raised by algorithms and artificial intelligence. Although their stance reveals concerns and an awareness of the risks, in principle they are not generally hostile to the use of algorithms and artificial intelligence-driven tools in our day-to-day lives, *as long as answers are forthcoming*.

The advantages mentioned during the various workshops held during the consultation day include personalised medical diagnoses, streamlined recruitment procedures which would thus become more neutral, simplified allocation of higher education places (APB) or the use of filters on online platforms to manage "the multitude of information". Many people view the new capacities for data analysis in a positive light: 63% thus believe that "sharing data for the common good" is worthwhile.

As the participants learned more during the consultation day, so their awareness of the risks grew: 32% considered that these "tend to be a source of error" at the end of the day, compared with 23% before the event. Whilst this might be a fairly modest rise by the end of a day devoted to the ethical issues, it coincided with a form of scepticism over the possibility of regulating algorithms in practice: "will the law be enough to ensure oversight across-the-board? Will we not still be having to resolve abusive practices after the fact?".

⁴⁹ http://www.huffingtonpost.fr/serge-tisseron/les-publicites-ciblees-cest-la-betise-assuree-interdisons-les_a_23220999/

Two founding principles for the development of algorithms and artificial intelligence: fairness and continued attention and vigilance

There are two distinct, albeit linked, intended outcomes to the discussions on algorithms and AI: the principles and the concrete means for putting these into practice.

Article 1 of the Data Protection Act states that “data processing shall be at the service of every citizen”. Today we need to be laying down the principles for achieving this general objective and for guaranteeing that artificial intelligence does serve humans – that it enhances us rather than claiming to replace us.

Do the principles enshrined in the Data Protection Act, of which we gave a reminder earlier on, still match with the issues that have been identified and to this general objective? Is there a need to promote new ones? Beyond the observation that these principles do not cover the full range of algorithmic and AI uses in practice, numerous requirements regarding algorithms raised in the public debate (fairness, accountability, intelligibility, explicability, transparency, etc.), is a sign of a sense of inadequacy, and perhaps even concern.

Here is a set of principles following on from the public debate. Two of these, bearing on fairness and continued attention and vigilance, stand out as particularly founding principles.

The principle of fairness

A principle formulated by the French Council of State

In its 2014 annual report on digital technology and fundamental rights, the Council of State outlined three recommendations calling for a “rethink of the principles underpinning the protection of fundamental rights”. The first of these concerned a principle of “informational self-determination”, guaranteeing data subjects control over the communication and use of their personal data. It has since been introduced into the Digital Republic Bill. The third had to do with the principle of “fairness” applied, not to all algorithms, but in a more restricted manner to “platforms”⁵⁰. According to the French Council of State,

“fairness consists of ensuring, in good faith, the search engine optimisation (SEO) or ranking service, without seeking to alter or manipulate it for purposes that are not in the users’ interest⁵¹”.

Platforms’ obligations towards their users in terms of the fairness principle as defined by the Council of State particularly include, on the one hand, the relevance of SEO and ranking criteria used by the platform with a view to providing users with the best possible service and, on the other, information about these criteria. The first obligation therefore limits the platform’s scope for establishing the algorithm’s criteria. The second makes the provision of information about the logic involved in the functioning of the algorithm an obligation on the part of the platform (so more than just a right that the user can choose to exercise or not).

This definition of fairness does not so much grant a right to users as it lays down an obligation with regard to controllers.

In a way, the beginnings of a principle of fairness can be found in the 1978 French Data Protection Act. For the right of information it upholds appears as a primary requirement in terms of fairness towards the data subject where an algorithm is processing his or her data. In addition to this, all data subjects shall have the right to obtain from the controller information about the logic involved in the functioning of the algorithm, and data subjects must consent to their data being processed: this is an obligation. That these rights are clearly stipulated in the 1978 Act means that such information must be provided “fairly” and that the algorithm must function along these lines.

French Council of State’s principle of fairness appears as particularly interesting as it mentions the notion of “users’ interest”. Indeed, it is not simply a question of the algorithm saying what it does and doing what it says: the principle of fairness also limits the extent to which the controller can determine the criteria by which the algorithm operates. Moreover, in the Data Protection Act, information is a right

⁵⁰ Il s’agissait de “soumettre [les plateformes] à une obligation de loyauté envers leurs utilisateurs (les non professionnels dans le cadre du droit de la consommation et les professionnels dans le cadre du droit de la concurrence)”. Les plateformes apparaissent comme des acteurs classant un contenu qu’il n’a pas lui-même mis en ligne.

⁵¹ *Le Numérique et les droits fondamentaux*, 2014, p.273 and 278-281

which may be exercised by the data subject by contacting the controller. With the principle of fairness, it is quite different since such information must be provided to the community of users from the outset⁵². So this is not about users' rights, but an obligation incumbent upon algorithmic platforms. In this respect, fairness could well represent a solution to the problem of unbalanced relations between controllers of algorithms and users.

The notion of fairness has also been on the agenda of additional talks led by the French Digital Council (CNUM). In its report entitled *Ambition numérique* (Digital Ambition, 2015), it made a proposal aimed at setting up an "algorithm fairness rating agency" which could rely on an open network of contributors. This would have a twofold objective. First, providing access, via a one-stop shop, to a whole series of information already collected by the various stakeholders as well as the existing tools. Second, creating a space for reporting problematic practices or malfunctions. This initiative could, in one form or another, provide several benefits on public knowledge of the issues, the balance between users and algorithmic platforms, the sharing of best practices between businesses and the detection by regulators of contested practices.

A principle to be extended to factor in the collective effects of algorithms

Regarding the definition given by the Council of State however, **it does seem timely to extend the principle, beyond platforms alone, to encompass all algorithms**⁵³. For example, should we not prevent an algorithm assisting doctors with medical decisions from using, or at the very least setting excessive store by, a criterion which could be appealing: optimising bed occupancy in a hospital?

In light of this, there would also be merits in the principle of fairness of algorithms being applied to algorithms or issues not dealt with by the legislation on personal data protection. In other words, it would apply to algorithms which do not carry out profiling of their users for the purposes of personalising their output (for example, to a search engine not displaying personalized results).

Lastly, we could consider the opportunity of **building on the Council of State's proposal by extending or, at the very least, clarifying the notion of "users' interest" such that not only the commercial and economic dimension of this interest is taken on board, but also its collective dimension**. This would entail considering that the algorithm's criteria must also not be completely at odds with certain key collective interests, particularly to do with the outcomes of personalization on collective benefits (as we previously highlighted). These collective interests can be

understood in two ways. On the one hand, we could be talking about the interest of categories put together through big data and algorithmic analysis (ad hoc groups formed by the cross-linking of certain characteristics), and which are likely to give rise to forms of discrimination. These categories are currently being discussed under the notion of "group privacy"⁵⁴. On the other, we could view this collective interest in terms of an entire society. For example, exposure to cultural diversity or diversity of opinions could be regarded as being in the "users' interest", who are not only consumers but also citizens and active members of a community (incidentally it would be advisable to refer specifically to "users' and citizens' interest").

The algorithm's criteria must not be completely at odds with certain key collective interests

With the rise of machine learning algorithms, the principle of fairness of algorithms – whilst evidently representing a solution as far as some major issues are concerned – comes up against a substantial stumbling block. As we have seen, these algorithms can behave in problematic ways for data subjects' rights, sometimes even without their designers knowing it (hidden bias and discrimination stemming from the correlations carried out by the system). The notion of fairness on the part of algorithm designers (which is what we are actually, usually, referring to when we speak of "fairness of algorithms") loses some of its scope the moment an algorithm behaves in a way which remains inscrutable even to its very designers. It must be possible either to speak of fairness of algorithms in the strict sense (but does this actually mean anything?), or to ensure that the algorithm will not behave in an undesirable way, even though we are not fully able to explain in principle what we mean by "undesirable". In other words, **a fair algorithm should not end up generating, replicating or aggravating any form of discrimination, even if this were to happen without its designers being aware**.

This last idea is thus much broader than the initial considerations raised earlier on the notion of fairness, which were focused on commercial and competitive concerns amid the development of decidedly unfair practices aimed at obtaining an advantage by manipulating the algorithm.

⁵² "Without ignoring trade secrecy, platforms would have to explain the general logic involved in their algorithms to users as well as, where applicable, the way in which users can change their settings."

⁵³ For the sake of settling any semantic quibbles, let us be clear that use of the expression "fairness of algorithms", rather than amounting to anthropomorphising a technical object (algorithm), is a handy shortcut for talking about the fairness of algorithm designers and processors.

⁵⁴ Brent Mittelstadt, From individual to group privacy in Big Data analytics, B. Philos. Technol. (2017) 30: 475. <https://doi.org/10.1007/s13347-017-0253-7>

The principle of continued attention and vigilance

While the principle of fairness appears to be a substantial founding principle, the one of continued attention and vigilance is more methodological, and must guide the way in which our societies model algorithmic systems.

One of the challenges identified has to do with **the changeable, scalable nature of machine learning algorithms**. This characteristic is compounded by the **unprecedented scale of the potential impact of algorithms** run by computer programs and therefore of the application of a single model. This increases the unpredictability and the likeliness of surprising outcomes. **How, then, should we tackle and regulate an unstable object**, which is likely to engender new effects as it grows and learns – effects that could not be foreseen at the outset?

Promoting a principle of “required continued attention and vigilance” could be a way to address this challenge, AI designers and operators taking into account this new characteristic. A further aim of this principle of required continued attention and vigilance would be to offset the phenomenon of excessive trust and weakened accountability which can arise in front of “black box” algorithms.

Finally, **this principle of continued attention and vigilance must have a collective** significance. More than algorithms, it is surely algorithmic systems that we should be speaking about – complex and long “algorithmic chains” made up of myriad stakeholders (developers, end users, companies that collect data for machine learning purposes, professionals who carry out this “learning process”, purchasers of a machine learning solution which they then intend to implement, etc.). This phenomenon – similar to the one that

can unfold along a subcontracting chain – plays a part in eroding the sense of responsibility, or simply awareness of the impacts that may result from such tools. For example, the data scientist may be placed at the very first stages of the algorithmic chain, he does not hold all the keys to understand the whole joint process. In the report it submitted to the CNIL, the Conseil National des Barreaux, the national institution that represents all practising lawyers in France, for its part drew attention to the fact that “the place where the program is implemented can have a very different understanding of ethics from the program designer”. What is more, inherent in data processing is the risk that excessive trust comes to be placed in a machine often perceived to be failproof and free from the bias that plagues human judgment and action. The Tricot Commission (which reflected on the French Data Protection Act) had already highlighted this risk back in the 1970s. Several of the speakers during the public debate also mentioned it this year. In all, the development of algorithmic systems is bringing with it a decrease in individual vigilance. In the face of the possible impacts of algorithms and artificial intelligence, there must be no question of allowing this type of indifference to grow. Collective continued attention and vigilance must be organised, with regard not only to known phenomena which we need to nip in the bud, but also to phenomena or impacts which could not necessarily be foreseen in the beginning but which could quite possibly come about because of the scale and changing nature of new algorithms.

**The development
of algorithmic
systems is bringing
with it a decrease
in individual vigilance**

Engineering principles: intelligibility, accountability, human intervention

Intelligibility, transparency, accountability

Given the opacity of algorithmic systems, **transparency** is an oft-cited requirement, with the idea that it could be a condition for fairness. According to the French Digital Council (CNNum), "first and foremost and in a general manner, this principle implies the transparency of the platform's behaviour, a prerequisite to ensure that what the service does in practice lives up to its stated promises. In relations between professionals, it applies to the pricing conditions for accessing the platforms and the conditions for opening up services to third parties⁵⁵". The opacity mentioned here concerns just as much the collection as the processing of data carried out by such systems, and therefore the role that they play in a certain number of decisions. Algorithms are not only opaque to their end users or data processors, however. With the rise of machine learning algorithms, the designers themselves are also steadily losing the ability to understand the logic behind the results produced. It is therefore at two levels that the issue of opacity must be addressed. Legal and procedural responses are necessary to create some necessary conditions for transparency, but technical ones are also needed.

Many consider the idea of transparency of algorithms to be too simplistic and ultimately unsatisfactory: transparency reduced to the simple publication of a source code would still leave the vast majority of the uninitiated general public in the dark about the underlying logic. Furthermore, at least where the private sector is concerned, the idea of transparency clashes with the right bearing on intellectual property. Algorithms are indeed likened to a trade secret which, if disclosed, could jeopardise an economic model.

Finally, companies can put forward good reasons for not revealing the source code or the criteria behind the functioning of an algorithm. Google, for example, is trying to make sure the results supplied by its search engine algorithm, PageRank, cannot be manipulated by stakeholders who would be able to turn its logic to their advantage.

Many specialists therefore recommend giving precedence to the requirement for algorithm explicability or **intelligibility** over transparency. What would seem to matter more than having direct access to the source code is the capacity to understand the general logic underpinning the way the algorithm works. It should be possible for everyone to understand this logic, which must therefore be explained in words rather than in lines of code. This is the opinion of Daniel Le Métayer, from the French Institute for Research in Computer Science and Automation (INRIA), for whom intelligibility entails probing the overall logic of the algorithm and specific results. It is also shared by Dominique Cardon: "What needs to be made transparent in the algorithm? Is it the statistical technique employed? Should the code be made visible? Even if there are merits to this, there are also reasons for its disclosure not to be an obligation. For example, in the search engine optimisation market, players are trying to sway the algorithm's output: this helps to understand one of the reasons why Google has not publicly disclosed its code. Making a computer transparent must above all involve an educational effort, in a bid to allow others to understand what it does. The key thing is not that the code be transparent, but that we understand what goes in and comes out of the algorithm and its objective. This is what must be transparent" (CNIL, public debate launch, 23 January 2017).

The idea of intelligibility (or explicability), in the same way as that of transparency, in any case ties in with the principle of fairness, as we might ultimately consider it to be a condition for the latter's implementation.

To end, introducing an obligation in terms of accountability or organising liability could be a way of addressing the phenomenon of diminishing accountability which algorithms and AI are tending to encourage. The idea would be that the roll-out of an algorithmic system systematically must give rise to a clear attribution of the liabilities that should be assumed in its operation.

Review the obligation for human intervention in algorithmic decision-making?

We have seen that the French Data Protection Act had established a principle banning any decision-making that produces legal effects on a data subject, if based solely on automated processing of personal data (in other words: based solely on the result provided by an algorithm analysing personal data). This principle is also set forth in the General Data Protection Regulation. However, this principle asserted in both these legislative texts is stripped of much of its substance because of very broad exceptions⁵⁶.

Moreover, courts seem to be invoking Article 10 of the 1978 Data Protection Act (which we are referring to here) less frequently now and interpretation of this article has tended to become less strict over the past forty years⁵⁷. An amendment to the Data Protection Act in 2004 has also facilitated automated decision-making, in the banking sector (credit scoring) for example. Although human intervention in this process is still a requirement, this takes the form of a right given to the data subject to ask, when he has been denied a credit, for the decision to be reviewed by a human. Human intervention then. But ex-post and only on request.

Without intending any value judgment, it seems possible to talk of a form of "drift" or shift in the threshold of society's tolerance for automated decision-making since the 1970s. The shift in the legal landscape and case law would seem to be a reflection of this ongoing change. As a result, should we not be reviewing the principle banning decision-making by a machine alone, where human intervention is therefore required? Reviewing it to accommodate new uses of AI, without giving it up altogether?

In its 2014 annual study, the Council of State stressed the need to ensure the effectiveness of human intervention. However, ensuring the effectiveness of human intervention for each decision may automatically imply preventing or limiting certain applications of algorithms and AI. Indeed, the purpose of automation is often to optimise or speed up a process by replacing humans. Genuinely effective human intervention as regards each decision thus risks having a dissuasive effect. We could in fact ask the question as follows: how can we get machines to perform tasks that were previously carried out by human intelligence (this is the definition of AI) without completely doing away with the need for humans? One solution is to consider the effectiveness of human intervention in other ways than at the scale of each individual decision. We could, for example, ensure that forms of human deliberation, where all sides can have their say, govern and guide the use of algorithms by examining and questioning the configuration, as well as all of the system's effects – direct and indirect. Rather than bearing on each individual decision, this process could thus, at intervals, concern series of decisions of varying number.

This would lead to the protection of freedoms being thought of more in collective than individual terms. We can also see how such a solution would tie in with the idea of an obligation for "continued attention and vigilance" mentioned earlier. This shift (from an individual interpretation to a collective interpretation of the obligation to ensure some form of human intervention in automated decision-making) could be more or less marked depending on the sensitivity of the applications in question and the risk/benefit ratio. In the health sector for example, should we consider the sensitivity of the stakes to outweigh the gains, which therefore could justify maintaining the obligation to guarantee human intervention for each decision?

How can we get machines to perform tasks that were previously carried out by human intelligence (this is the definition of AI) without completely doing away with the need for humans?

⁵⁶ On this point in the GDPR, see for example: Wachter, Sandra, Brent Mittelstadt, & Luciano Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". Social Science Research Network, December 2016

⁵⁷ See for example the CNIL's deliberation on the GAMIN draft ruling, in 1981: the CNIL rejected this Ministry of Health-led draft. Even the guarantees that the Ministry provided to ensure effective human intervention in the detection of minors at risk of mental health problems, at issue here, were rejected. We could nevertheless wonder, on reviewing the case, whether the CNIL would adopt the same position today, at a time when we seem to have come round somewhat to the idea of seeing algorithms feature ever more prominently in ever higher-stake areas. For example, whilst the decision to eliminate candidates solely on the basis of automated processing does not seem so far-fetched as to constitute science-fiction, it is unlikely that many people in our society are ready to accept it.

From principles to policy recommendations

How can the aforementioned principles be put effectively into practice? The following pages list the main recommendations that came to light following the public debate that the CNIL organised from January to October 2017, rounded off by the consultation of reports that various institutions in France and abroad have already submitted (including the OPECST, CERNA, CNNum, Council of State, CGE, the White House, France IA, INRIA and AI Now).

A general point that came to the fore during the discussions is that the solutions call for a diverse range of actions on the part of various stakeholders (algorithm designers, professionals, businesses, the public authorities, civil society and end users). AI and algorithmic

systems are complex socio-technical objects, handled by long and complex chains of stakeholders. **Action is therefore required right across the algorithmic chain** (from the designer to the end user, via the system trainers and operators), **through both technical and organisational approaches**. Algorithms are everywhere, and they are therefore everyone's business.

The law should not be the only lever: the solution is to be found by rallying all of the stakeholders to the cause. A certain number of the recommendations outlined below do not specify, incidentally, whether precedence should be given for their implementation either to the law or to voluntary initiatives of various stakeholders.



WHAT THE PUBLIC THINK

Participants in the public consultation which the CNIL organised in Montpellier on 14 October 2017 (see "Organisation of the public debate on the ethical issues of algorithms and artificial intelligence") came up with recommendations. These largely overlap with the ones that were formulated at other times during the public debate.

- The desire that humans maintain control over the development of algorithms appears paramount (95% in favour), with an excessive delegation of decisions to algorithms and AI having been deemed harmful. The participants' view supports the aforementioned idea of a principle of continued attention and vigilance: 97% would like to "retain the human dimension, retain a dose of subjectiveness and not disengage completely" and 91% believe that "the user should play the role of a learner each time s/he uses an algorithm, so as to grasp its limits and be demanding of developers whenever necessary". In medicine for example, some citizens reckon that certain decisions should always be discussed in a collegiate manner.
- Improving the training of algorithm designers to certain ethical matters is an option that arose out of several workshops and garnered almost complete consensus: 97% of participants think that "developers should build a certain ethical framework into their practices and resist some tempting market incentives". 94% thus call for the development of ethical charters in this regard and 56% would like social and human science experts to help developers to better gauge the impact of their work on society. Training also concerns algorithm users: 82% of participants are in favour of a mandatory continuing education for doctors who use decision support systems. More generally, individuals want to know and understand, thus asking for a more lifelong learning on the subject of digital technology. In order to address problems of inequality associated with such objects, all citizens are unanimous in urging "informal education on the subject of digital technology" and the "development of school syllabuses for digital "literacy" in terms of both the object and the issues".



WHAT THE PUBLIC THINK (cont.)

- Each group also strongly underscored the need for enhanced rights in terms of information, transparency and explanation regarding the logic involved in the way the algorithm works. At first, the participants seem to be demanding the possibility of being notified each time an algorithm is being deployed: 88% of them are of the opinion that an employer who uses an algorithm must inform candidates thereof without fail. 78% of them are in favour of source codes being made publicly available, even though this is considered insufficient to understand the results produced by an algorithm. In the case of the university admissions online portal, "APB", for example, 78% of participants call for a greater practical guidance so as to understand the ins and outs of its use; 85% see users' feedback as an invaluable resource for "improving the user-friendliness of the procedure". Further, when an algorithm's criterion is grounded in political choices (drawing lots for example), this should not be concealed but, quite the opposite, made explicit and understandable (according to 94% of participants). Note that although there is a desire for transparency, it is not unanimous and, what is more, some people are aware, if not convinced, that it may not be enough.
- An overwhelming majority of participants call for regulatory efforts on the part of the State to identify bias, "prevent malpractice, draw up statistics and make approvals compulsory" (97%). Many recommend setting up an independent body to perform scientific testing on algorithms, "in the same way it applies to medicines before being offered for sale on the market" (84%). In the long term, checking at regular intervals that the algorithm "still meets the set objectives" is another idea that was floated during the debates (63% are in favour). There is also strong support (94%) for new laws to improve the extent to which ethics is taken on board in law "through codes of conduct and charters, training and dialogue".
- Some participants also highlighted the importance that civil society organise itself so as to be more prepared for these new technological objects. For instance have been discussed: the role of associations (patients' associations in healthcare for example), whistleblowers' protection, or the support lent to alternative networks to the online platforms whose algorithms raise questions.
- Lastly, the discussions revealed a strong attachment to personal data protection and privacy. The question of who owns our data and what uses are made of it was deemed pressing among some work groups, on the health theme in particular but also as regards employment (concern over the possibility that algorithms analyse data that would be collected outside the company).

RECOMMANDATION 1

Fostering education of all players involved in the "algorithmic chain" (designers, professionals, citizens) in the subject of ethics

Training for the general public

Citizens have a key role to play in algorithmic systems. On the one hand because algorithms are having an increasing impact on their lives and, on the other, because they are particularly well placed for spotting any abuses. Enabling

them to understand these new technologies so that they can use them in a confident, active and informed way is a requirement. It also corresponds to a demand they themselves have made – as underscored during the public consultation organised in Montpellier by the CNIL on 14 October 2017.

There is broad consensus over the need to develop a "new digital literacy" from primary school right through to university level. The CNNum is just one of the stakeholders already making such a case⁵⁸. This digital literacy would obviously include a basic familiarity with algorithms, the grounding in which could, incidentally, be given very early on (via exercises which do not necessarily involve using digital devices).

⁵⁸ <https://cnnumerique.fr/education-2/>

Encouraging digital mediation initiatives in local areas can also contribute to make the population widely familiar with algorithms. In other words, a form of public digital education including a basic grounding in data and algorithms. Examples of such initiatives include FING (Info Lab), La Péniche in Grenoble (Coop-Infolab), and POP School in Lille.

Training for algorithm designers

Algorithm designers (developers, programmers, coders, data scientists, engineers) occupy the first link in the algorithmic chain. This represents a particularly high-stakes stage. The technical complexity of their jobs is, moreover, likely to make their actions opaque (and therefore difficult to supervise) to other players. It is paramount that they have the fullest possible awareness of the ethical and social implications of their work and of the very fact that these can even extend to societal choices which they should not by rights be able to judge alone. And yet **the way the workplace and economy are organised in practice tends to create a silo mentality. Tasks are assigned to separate departments, each of them being likely to ignore the implications of their activity outside their own silo.** What this means is that training is a first essential step for algorithm designers to be able to grasp the sometimes very indirect implications of their action both for individuals and society, thus making them aware of their responsibilities by learning to show *continued attention and vigilance*.

In this regard, there could be merits to including the social and human sciences approach (sociology, anthropology, management, history of science and technology, information and communication sciences, philosophy and ethics) to these issues in engineer and data scientist training.

The development of these courses could benefit from the inclusion of social and human sciences and technical approaches within interdisciplinary laboratories.

There are already some initiatives under way in this context. We could mention the ENSC (Bordeaux's Cognitique Institute), a prestigious graduate school which includes the social and human sciences in its engineers' training programmes, and Costech (which stands for "Knowledge, organisation and technical systems"), at the Université Technique de Compiègne (UTC).

Finally, if we are to guarantee that artificial intelligence does not foster forms of ethnocentrism, it is vital to encourage cultural, social and gender diversification in the occupations involved in designing algorithms.

The first step to getting more women working in these specialities particularly entails efforts to increase incen-

tives for female students to access more widely training programmes.

Training for professionals who use algorithms

To make sure that no links along the algorithm chain deployment chain are overlooked, it is also necessary to provide training for professionals who are required to use such systems as part of their jobs. This would particularly involve forearming them against the risk of a diminished sense of responsibility and loss of autonomy that can result from using tools which sometimes work like black boxes whose effectiveness cannot be questioned. It is crucial to guard against excessive trust by raising awareness of the ethical dimensions of a decision-making process that must not exclude human intervention and by honing critical thinking in some particularly sensitive sectors, such as medicine, recruitment, justice and perhaps now marketing above all, where the antisemitic categories recently generated by Facebook's machine learning algorithms are a stark wakeup call to the sharpness of the risks. This training should particularly include, in a multidisciplinary mindset, consideration of the specific issues that these tools raise in each sector. A doctor who uses an AI-based diagnosis support system, for instance, should be made explicitly aware of the possible development of bias. He should also be capable of understanding the implications of the tool he is handling and the consequences of any mistakes.

One option could thus be to create a sort of "licence to use algorithms and AI" in some sectors, which could be earned through specific training modules administered by specialist schools and universities.

Raising the awareness of public stakeholders about the need for a balanced and "symmetrical" use of algorithms

Similarly, it would be advisable to educate public stakeholders in the need for a balanced and symmetrical deployment of algorithms. At a time when the latter are being increasingly rolled out to crack down on fraud and for carrying out checks, we should avoid the public to reach the mistaken conclusion that they can only be used for monitoring and law enforcement purposes (which are actually useful to individuals themselves). The risk would be a form of distrust to thrive, that would ultimately undermine the deployment of algorithms and the harnessing of their benefits. Administrative and political leaders must therefore be convinced of the merits of tapping into the potential offered up by algorithms, of which the benefits to individuals are immediately apparent and which help to improve access to rights (detection of non-take-up of social benefits)⁵⁹.

⁵⁹ In an assessment of public policies fostering access to social rights, in 2016 MPs Gisèle Biémouret and Jean-Louis Costes suggested "using anti-fraud tools to try and reduce non-take-up of social rights". See: Information report of the public policy assessment and oversight committee on the assessment of public policies in favour of access to social rights.

RECOMMENDATION 2

Make algorithmic systems understandable by strengthening existing rights and organising mediation with users

The law give us first clues to how we can address the opacity, for individuals, of the algorithms that profile them and the logic they follow (to grant them a bank loan for example). As we have seen, some provisions pave the way to an initial form of intelligibility and transparency for many years already⁶⁰.

However, many analyses agree that these provisions are insufficient to effectively demystify algorithm-driven systems and ensure intelligibility, transparency and fairness. One way to tackle this challenge would be to bind system controllers to an obligation (rather than merely where requested by the data subjects) to communicate information in a clear and understandable manner enabling the logic involved in an algorithm to be grasped. This has already been provided for, in fact, in the French Digital Republic Bill for algorithms used by public authorities⁶¹.

It also seems entirely relevant for this requirement (whether determined by the law or freely adopted by the stakeholders) to concern algorithms that do not process the personal data of their users too, insofar as they are likely to have significant collective impacts – regardless of the fact that these are not direct impacts on individuals themselves (see, in particular, “The limits of the current legal framework” and “The principle of fairness”).

Such an obligation, enshrined in the law, could be usefully extended by private initiatives setting a virtuous cycle in motion. In the case of stakeholders with websites on which data subjects have an account they can log in to, could be made available: information about their “profile”, or the data processed and inferred and the logic underpinning the way the algorithm works. In this way data subjects could correct and update their profile and personal data easily.

This updated legal framework could go hand-in-hand with the development of best practices by the stakeholders, with the help of soft law instruments.

The problem of the opacity of algorithms also stems from the fact that **algorithmic system controllers are not, in the vast majority of cases, reachable or accessible in practice to get hold of information and explanations.** This is also tied in with a lack of accountability of such systems, where users find it impossible to hold anyone to account. **It is therefore necessary to organise a form of “reachability” of algorithmic systems,** particularly by systematically identifying within each company or authority a team that is responsible for an algorithm’s operation the moment this processes the data of humans. Deliberate and clear communication of the identity and contact details of this person or team is also necessary to ensure they can easily be contacted and they have the means to respond swiftly to the requests received.

In addition to reachability, committed efforts must also be made to organise mediation and dialogue between systems and society, along the lines of the ideas developed by the Fondation Internet Nouvelle Génération (FING) as part of the initiative “NosSystèmes”. For FING has found that “reaching the technical controller is not enough”. Accordingly, it suggests setting up teams dedicated to the quality of user dialogue and a “mediation percentage”. While algorithms allow for economies of scale, factoring in the percentage of a project’s budget devoted to mediation efforts (setup of visualisation tools, mediation team, partnership, checks that information has been properly understood, etc.) could – via certification procedures – be a way to enhance and bestow a competitive edge (in terms of image in the eyes of consumers) upon virtuous systems.

RECOMMENDATION 3

Improve the design of algorithmic systems in the interests of human freedom

More than the algorithm alone, or even the program running the algorithm, it is to the whole algorithmic system that we need to be turning our attention to understand and monitor its effects. Many recent discussions stress the importance of taking the design of algorithmic systems into account – i.e. the interface between the machine and its user.

⁶⁰ Not least Article 39 of the Data Protection Act, organising the right of access.

⁶¹ Article 14.1a of the General Data Protection Regulation also provides for such information.

What this means is working on their design to counter the “black-box-like” nature that algorithms can assume. They do so by coming across as inscrutable systems displaying results without putting their own limits into perspective or explaining the way in which they are built. Algorithms also do so by giving off an air of prestige on account of the neutrality and infallibility with which we are so quick to credit machines.

So we need, instead, to be promoting a design conducive to empowering individuals with more autonomy and scope to think; a design to righting the imbalance that algorithms can create to our detriment; overall a design that enables us to make clear and informed decisions.

For example, setting up visualisation systems handing back more control to users, by providing them with better information, would be one way of doing this. **With visualisation tools, users can understand why they have been given recommendations or, better still, receive more appropriate recommendations in return.** In this way, they are placed on an active footing. The aim is to give individuals a handle on the criteria (or at least part of it) determining the response provided by the algorithm; perhaps even enabling them to test out different responses according to different configurations. One example of a visualisation tool was provided during the public debate with the presentation of the “Politoscope”⁶². Developed by the Complex Systems Institute of Paris-Île de France (ISC-PIF), the Politoscope gives the general public an insight into masses of data and the activity and strategy of political communities on social media, Twitter in particular. By lifting the lid on the practice of astroturfing, it helps to offset it: this is the attempt

by highly organised groups to manipulate social media and thereby push certain themes to the top of the national political agenda. In this way, the Politoscope is helping to restore balance in algorithm use, with a view to safeguarding the democratic access to information.

Through design, the whole relationship between humans and machines can be adjusted, to empower us and increase our ability to make informed decisions – as opposed to taking this capacity to make choices away from us and giving it to machines. In short, this is about giving substance to the principle of continued attention and vigilance that we discussed above.

The concept of “testability” recently suggested by FING as part of its “NosSystèmes” expedition could also represent a principle governing the design of user-friendly virtuous algorithmic systems that allow users full scope to act. This concept is about enabling users to “test out” the systems by playing around with their settings. For example, this could have meant giving users of the university admissions portal, “APB”, the opportunity to perform a “practice run” by seeing what results are given for different choices before entering their final choices. We could thus imagine an online search engine giving its users the option of running several searches according to different criteria. The idea behind testability is that **having a go ourselves is the key to direct understanding** – much more, arguably, than access to a source code which would be indecipherable to the vast majority of us.

RECOMMENDATION 4

Set up a national platform for auditing algorithms

Developing algorithmic auditing to check their compliance with the law and fairness is often billed as a solution to ensure their fairness, accountability and, more broadly, their compliance with the law.

Through design,
the whole relationship
between humans and machines
can be adjusted,
to empower us and increase
our ability to make
informed decisions

⁶²<https://politoscope.org/>

This would first require building the capabilities of public authorities in this regard. Algorithmic auditing is not a new concept. In 2015, the Enforcement Committee of the "Autorité des marchés financiers" (which regulates participants and products in France's financial markets) examined the "Soap" algorithm in handing down its decision dated 4 December 2015 regarding the firms Euronext Paris SA and Virtu Financial Europe Ltd. Similarly, for its inspecting activity the CNIL draws on the expertise of information system auditors. The Autorité de la concurrence (an independent administrative authority specialising in supervising anticompetitive practices) must also increasingly ground its work in its ability to audit algorithms.

It is therefore essential that public authorities do their utmost to open up the source code of deterministic algorithms – at a time when the means at their disposal are increasingly falling short of what the surge in demand requires. Other sector-specific regulators with no auditing capabilities are thus calling on the CNIL to intervene on their behalf. **Today we therefore need to identify what resources the State has available, as well as the different needs, and pool the expertise and means to hand within a national platform.**

Such a platform should also tackle the challenge raised by the development of machine learning. This is prompting some people to point out that the solution of examining source codes is hardly realistic when there are millions of lines of code to be analysed. Now, auditing does not necessarily mean opening up source codes. It may also take the form of ex-post checks of the results produced by algorithms, or tests using simulated profiles for example. Significant research efforts should be geared towards these auditing techniques, which are based on reverse engineering (see the next recommendation).

In practice, these audits could be performed by a public body of algorithm experts who would monitor and test algorithms (by checking that they do not practise discrimination for example). Given the size of the sector to be audited, another solution could involve the public authorities accrediting private audit firms on the basis of a frame of reference. Some private initiatives are already up and running. Cathy O'Neil, who we have already cited several times in our report, has set up the company "O'Neil Risk Consulting & Algorithmic Auditing", whose mission is to help companies to identify and correct the risks of the algorithms they use.

Quite separately from a requirement to perform an auditing procedure, companies and public authorities would be well

advised to adopt certification-type solutions. These certifications could fuel a virtuous momentum. On the one hand, they would guarantee that algorithms practise fairness and non-discrimination. On the other, they would shed light on existing efforts to set up a design as well as proactive and appropriate Information (in keeping with the recommendations above) going beyond the strict legal obligations.

RECOMMENDATION 5

Increasing incentives for research on ethical AI and launching a participatory national worthy cause on a general interest research project

Encourage explanations on the functioning and logic of algorithms

Research policies should focus increasingly on providing regulators, businesses and citizens with robust tools for checking, controlling and monitoring the impacts of algorithms and artificial intelligence, and unpicking the logic behind them.

It would be well worth channelling major investment towards the development of **reverse engineering techniques** to "test" the non-discriminatory nature of algorithms and AI, the ability to **pre-process data to reduce the risks of discrimination** by pinpointing and clearing up bias in training datasets⁶³ and **the generation, by algorithmic machines using machine learning, of explanations in natural language of their output.**

In France, the TransAlgo project led by INRIA (French Institute for Research in Computer Science and Automation) is already seeking to galvanise action on these issues by developing a scientific platform. The Algodiv project (algorithmic recommendation and diversity of information on the web), meanwhile, sets out to provide answers to the questions raised by filter bubbles: are algorithms harming diversity and serendipity? In a nutshell, what these projects are setting out to do is shed more light on a certain number of issues discussed in this report.

Initiatives which combine interdisciplinarity, cutting-edge research and tool development should be supported in France, along the lines of the initiative led by Professor Katharina Anna Zweig in Germany who, in 2017, set up the Algorithmic Accountability Lab. In addition to drawing on the hard sciences, technical sciences and human sciences

⁶³ Construction of a non-biased dataset was the focus this year of a project led by the Open Law association, a partner in the public debate organised by the CNIL.

to perform analyses (in line with the idea that algorithmic systems can only be understood, predicted and monitored in the context of their application), this lab is working on developing a transparent, ethical and accountable design of algorithmic decision-making systems (ADM). It also offers pedagogical tools concerning the risks and promises of ADM⁶⁴ for the mainstream public and decision makers alike.⁶⁵

Another example is the recent creation in the United States of the research institute AI Now (within New York University), which examines the social implications of artificial intelligence. The involvement in the institute of the "Partnership on AI" consortium, whose founding partners particularly include Amazon, Apple, Google, IBM and Microsoft, nevertheless highlights the close attention that should be paid to the membership of such institutes. As recently pointed out by former academic Cathy O'Neil, the importance of bringing researchers on board in shedding light on the social impacts of AI is associated with the freedom of inquiry that academics enjoy⁶⁶.

Develop research infrastructure that respects personal data

The development of data-friendly AI is becoming increasingly important at a time when citizens in Europe, but to a broader extent worldwide also, are more and more concerned about the protection of their personal data and the risks generated. Various solutions can be put forward with a view to forging a new balance based simultaneously on the strengthening of researchers' access to substantial datasets, and of the security of this data.

To begin with, it implies the development of secure spaces for accessing data for the purposes of research and training AI algorithms. Work along the lines of what the OPAL project undertakes could contribute towards this goal. This project is aimed at building an infrastructure on which telecom operators' data is stored and can be analysed in complete safety, by certified open algorithms made available to users and which can be broadened by the community. With such systems, the data is not directly accessible to the people processing it, thereby guaranteeing the protection of data subjects. Certification of the algorithms that can be used to analyse these datasets has an ethical data filtering function, which particularly makes it possible to tackle the challenges posed in terms of "group privacy"⁶⁷.

Databases accessible to public stakeholders, such as the CASD (Secure Access Data Centre), used in France to make public authority databases available for research purposes, are also a solution to be delved deeper into.

Launch a participatory national worthy cause to boost research in AI

The ability to access huge volumes of data forms one of the cornerstones of AI research development. Contrary to common belief, French and European legislation provides a sufficiently open framework for supporting ambitious industrial policy and research in this regard. Over and above the possibilities we have already outlined, the creation by the General Data Protection Regulation (GDPR) of a "right to data portability", which enables data subjects to retrieve their personal data in the possession of private stakeholders, paves the way to major opportunities of which we still have little idea for the most part.

The public authorities could act as driving forces in bringing the latter to fruition. For instance, they could launch a national worthy cause or a key research project based on data contributed by citizens exercising their right to data portability with private stakeholders and transferring their data for the benefit of a project in the general interest. The State would guarantee that the project respects freedoms and could, for example, back the creation of a management chart (modelled on FING's "NosSystèmes" project) for use by data subjects. In this way, the public authorities would start to leverage the vast potential offered up by the creation of this right – with repercussions beyond this one project.

Private stakeholders could naturally bring their own datasets to the table and thus play a part in this national worthy cause.

RECOMMENDATION 6

Strengthen ethics within businesses


What has become clear today is that businesses are also being called to identify any irregularities or adverse effects before algorithms with far-reaching impacts are deployed. They are also expected to keep a constant watch out for emerging problems, whether imperceptible or unnoticed at the outset, by providing a counterpoint to the operational perspective. The aim is also to gain an overview of algorithmic chains given their tendency, as we have highlighted, to divide tasks and concerns into separate compartments. Following the same mindset, there is a need to organise forms of dialogue between practitioners, specialists outside the company, stakeholders and communities involved in the use of algorithms alongside researchers in the social and human sciences.

⁶⁴ Algorithmic Decision Making Systems.

⁶⁵ This laboratory has already completed several projects, including the "data donation" project ("Datenspende Projekte" in German; <https://datenspende.algorithmwatch.org/>), during which more than 4,000 users observed Google's search results on the 16 main candidates over the months running up to the German parliamentary election. The underlying idea was to measure the impact of Google's personalisation of search results so as to confirm or invalidate the "filter bubble" theory.

⁶⁶ <https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html>

⁶⁷ Previous projects have shown that the use of anonymised data was likely to lead to problematic uses from an ethical point of view (targeting of population groups – and not necessarily individuals – on an ethnic basis in contexts of conflict, or actuarial segmentation, among others).




What has become clear today is that businesses are also being called to keep a constant watch out for emerging problems, whether imperceptible or unnoticed at the outset

Several approaches to putting this requirement into practice could be considered.

One solution could entail setting up ethics committees within companies that use algorithms with far-reaching impacts. The makeup and working procedure of such committees are key considerations. Whether or not their reports, and list of members, are published, the possible degree of independence: there is a wide range of possibilities.

Assigning this task to the Corporate Social Responsibility role or to professional ethics specialists could also be an option. Coordination of ethical discussions in the private sector could also be ensured through networks formed by industries or sectors, so that best practices can be disseminated and emerging problems detected early on. Another idea might involve sector-specific ethics committees organising a form of ethics watch, instead of having committees set up within each company – even though the guarantees in this respect would not be as strong.



This networking should set out to compile and keep up-to-date sector-specific ethical frameworks (such as ethical charters, codes of professional conduct or codes of ethics for example) and revise pre-existing codes of professional conduct so as to take the introduction of algorithms and AI systems into account.

Such discussions ought, in return, to lead to the addition, to companies' codes of professional conduct, of a chapter on the issues raised by algorithms (for example by clarifying where to draw the line when designing system settings, obligations bearing on the quality and updating of datasets used to train algorithms and so on).

Our intention in outlining the various possibilities in the paragraphs above is to show that it would certainly be worth conducting specific debates on the exact blueprint for how best to proceed. There can evidently be several different views in this regard.

CONCLUSION

The principles and recommendations set out at the end of this report are the result of the summary that the CNIL drew up of the discussions held during the national public debate it oversaw from January to October 2017, with help from sixty partners.

The policy recommendations have been expressed very broadly, calling on the fullest possible spectrum of public and private stakeholders. In light of the challenges raised by algorithms, the whole of civil society (citizens, businesses and associations) must get involved, pay attention and ask questions as they navigate a complex world. The intention was not, therefore, to state that the law could be the only appropriate means for applying them. Quite the opposite: most of the recommendations may be interpreted as able to achieve concrete expression through either a binding legal framework or voluntary adoption on the part of stakeholders – with solutions embracing varying degrees of these two extremes being possible.

The discussions brought two founding principles to the fore, and since some of the main ethical challenges raised by artificial intelligence can be subsumed under them, they merit particular attention.

First, **the substantial principle of fairness of algorithms**, which builds on the principle already proposed by the French Council of State (see the section “The principle of fairness”). This version factors in the idea of fairness towards users, not only as consumers but also as citizens, and even towards communities whose lifestyles could be affected by algorithms, whether or not these process personal data.

Second, **a more methodological principle is that of continued attention and vigilance**. This is to be understood not as a vague incantation but as a substantiated response to three central challenges facing the digital society. One, the changing and unpredictable nature of algorithms in the “machine learning age”. Two, the

silo mentality affecting the organisation of algorithmic chains, which leads to action being carried out in isolation, indifference to the overall impacts of the algorithmic system and diminishing accountability. Three, the risk of excessive trust being placed in machines, which a form of human cognitive bias leads us to consider as being-fail-proof and free from bias. The principle of continued attention and vigilance is basically aimed at organising the ongoing state of alert that our societies need to adopt as regards the complex and changing socio-technical objects that algorithmic chains or systems represent. This state of alert means constantly subjecting to scrutiny, to methodical doubt. This first and foremost concerns individuals, who form the links of algorithmic chains: they need to be given the means to be on the look-out, in an aware and active manner, always seeking answers, in this digital society. But it also concerns the other key players in our society: businesses of course, to model virtuous algorithmic systems, as well as others.

Owing to the universal approach through which they came about, these principles could form part of a new generation of principles and human rights in the digital age: a generation which, after those of rights-freedoms, property rights and social rights, would be that of “system-rights” organising the dimension underpinning our digital world. Are they not destined to be upheld as general principles for the global governance of Internet infrastructure? At a time when France and Europe are setting out their positions regarding artificial intelligence, the question is an entirely relevant one.

The principles of fairness and continued attention and vigilance could form part of a new generation of principles and human rights in the digital age: system-rights organising the dimension underpinning our digital world

ACKNOWLEDGEMENTS

The CNIL is sincerely grateful to the individuals and institutions that took part in this collective thought process.

The partners in the public debate

- Allistene's research committee on ethics (CERNA)
- Bordeaux's Cognitive Institute (ENSC)
- Bordeaux University
- Caisse des dépôts et consignations (CDC)
- Club des Juristes (thinktank)
- Collège des Bernardins
- Complex Systems Institute of Paris Ile-de-France (ISC-PIF)
- Confédération française de l'encadrement – Confédération générale des cadres (CFE-CGC, trade union)
- Communication Publique
- Conseil National des Barreaux (national institution that represents all practising lawyers in France/CNB)
- Conseil Supérieur de l'Audiovisuel (independent authority to protect audiovisual communication freedom/CSA)
- Conservatoire National des Arts et Métiers (leading higher education and research institution dedicated to adult continuing education/CNAM)
- Douai court of appeal
- ESCP Europe, IoT Chair
- Etalab (body that works in France on data sharing in the public sector)
- "Familles rurales" association
- Federal University of Toulouse
- French Association for Artificial intelligence (AFIA)
- French Association for Employment Law (AFDT)
- French Development Agency (AFD)
- French governmental advisory council on bioethics issues (CCNE)
- French Insurance Federation (FFA)
- French National Center for Scientific Research (CNRS)'s ethics committee (COMETS)
- FO-Cadres (trade union)
- Fondation Internet Nouvelle Génération (FING)
- Fotonower
- Génotoul societal (bioscience and ethics platform)
- Groupe VYV (MGEN – ISTYA – Harmonie)
- Imagine Institute on genetic diseases
- INNOvation Ouverte par Ordinateur (INNOOO)
- Institut Mines-Télécom (IMT), Research Chair "Values and Politics of Personal Information"
- Laboratory for Collective and Artificial Intelligence (LICA)
- Law Department of Université Catholique de Lille, Centre of research on relations between risk and law
- Law Department of Université Catholique de Lyon
- Ligue des Droits de l'Homme (Human Rights League/LDH)
- Ligue de l'Enseignement (Education League)
- Lille 2 University
- Lille Association of Lawyers
- Lyon's administrative court of appeal
- Microsoft
- Ministry of Culture, via the General Directorate of Media and Cultural Industries (DGMIC)
- Ministry of National Education, via the Directorate of Digital Technology for Education (DNE) and its Numéri'lab
- National Academy of Technologies of France
- National Institute of Higher Studies on Defence (IHEDN)
- National Institute of Higher Studies on Security and Justice (INHESJ)
- National Institute of Applied Sciences (INSA)
- Necker Hospital
- OpenLaw (association)
- Paris II University
- Randstad
- Research Centre of the National Gendarmerie School of Officers (CREOGN)
- Rhône Département-level Council of the Medical Association
- Renaissance Numérique (thinktank)
- School of Advanced Studies in the Social Sciences (EHESS)
- Sciences Po Lille
- Sciences Po Paris
- Société informatique de France (association devoted to computer science/SIF)
- The Future Society at Harvard Kennedy School, AI Initiative
- Universcience
- Visions d'Europe (association)

The other contributors

- Arbre des connaissances (association)
- Autorité de contrôle prudentiel et de résolution (French authority responsible for the supervision of the banking and insurance sectors/ACPR)
- Autorité des marchés financiers (authority which regulates participants and products in France's financial markets/AMF)
- Montpellier Méditerranée Métropole and its President, Philippe Saurel
- City of Montpellier

The 37 citizens who took part in the public consultation organised in Montpellier on 14 October 2017.

Jérôme BERANGER • Nozha BOUJEMAA •
Dominique CARDON • Jean-Philippe DESBIOLLES •
Paul DUAN • Flora FISCHER • Antoine GARAPON •
Roger-François GAUTHIER • Hubert GUILLAUD •
Rand HINDI • Jacques LUCAS •
Camille PALOQUE-BERGES • Bruno PATINO •
Antoinette ROUVROY • Cécile WENDLING

LIST OF EVENTS ORGANISED FOR THE PUBLIC DEBATE

From the end of March until the beginning of October, the CNIL oversaw and coordinated 45 events on algorithms and artificial intelligence. Some of the initiatives were designed specifically for the public debate launch, while others were part of projects already being led by various stakeholders – public institutions, associations, research centres – for which these issues were already striking a note of concern.

Many stakeholders chose to broach algorithms in a specific sector (healthcare, employment or education for example), while others took an overall approach to this technological subject matter. Expert workshops for a limited audience and mainstream events for the general public (citizens and students for example) alike were held throughout the process.

More information on the events can be found on the CNIL's website.

- 23/01/2017 ■ **LAUNCH EVENT**
ROUNDTABLE SESSIONS "Algorithms and humans"
and "Fairness, transparency and plurality of algorithms"
> **CNIL**
- 23/03/2017 ■ **SYMPOSIUM** "Towards new forms of humanity?"
25/03/2017 > **Universcience**
- 31/03/2017 ■ **CONFERENCE** "Algorithms and law"
> **Lille II University**
- 06/04/2017 ■ **CONFERENCE** "The choice in the age of Big Data"
> **Sciences Po Lille and Visions d'Europe**
- 08/04/2017 ■ **DEBATE** "The governance of emerging technosciences"
> **German American Conference at Harvard University**
- 18/04/2017 ■ **DEBATE** "Transatlantic perspectives on: AI in the age of social media; privacy,
security and the future of political campaigning"
> **The Future Society at Harvard Kennedy School**
- 18/04/2017 ■ **ROUNDTABLE SESSIONS** "Big Data, human resources: algorithms on the agenda"
> **FO-Cadres**
- 04/05/2017 ■ **CONFERENCE** "Fairness of algorithmic decision-making"
> **Toulouse III – Paul Sabatier University**
- 16/05/2017 ■ **DEBATE** "Will digital technology spell the end of the rule of law?"
> **Collège des Bernardins**
- 19/05/2017 ■ **SYMPOSIUM** "Predictive justice"
> **Douai Court of Appeal, Lille Association of Lawyers and Law Department of Université Catholique de Lille**
- 02/06/2017 ■ **WORKSHOPS** "Fairness of algorithmic decision-making and processing"
> **LabEx International Centre for Mathematics and Computer Science in Toulouse**

- 08/06/2017 ■ **DEBATE** "Algorithms in healthcare: what ethics?"
> **Groupe VYV (MGEN – ISTYA – Harmonie)**
- 14/06/2017 ■ **ROUNDTABLE SESSION** "Artificial intelligence: ethics, at the intersection of HR and Big Data"
> **Confédération française de l'encadrement – Confédération générale des cadres (CFE-CGC)**
- 16/06/2017 ■ **DEBATE** "Algorithms, employment and ethics"
> **French Association for Employment Law (AFDT)**
- 19/06/2017 ■ **DAY** "Ethical algorithms, a moral requirement and competitive advantage"
> **Allistene's CERNA and Société Informatique de France (SIF)**
- 19/06/2017 ■ **SYMPOSIUM** "Human, non-human in the age of artificial intelligence"
> **Paris II University**
- 21/06/2017 ■ **SYMPOSIUM** "Artificial intelligence: autonomy, delegation and accountability"
> **Bordeaux's Cognitive Institute (ENSC)**
- 22/06/2017 ■ **WORKSHOP** "Ethics of algorithms: implications for healthcare"
> **Genotoul (bioscience and ethics platform)**
- 22/06/2017 ■ **CROWDSOURCING WORKSHOP** "Artificial intelligence and law"
> **OpenLaw**
- 22/06/2017 ■ **SYMPOSIUM** " The many dimensions of data "
- 23/06/2017 ■ > **Institut Mines-Télécom, Values and Politics of Personal Information Research Chair**
- 27/06/2017 ■ **SYMPOSIUM** "Security and justice, the challenge of the algorithm"
> **National Institute of Higher Studies of Security and Justice (INHESJ)**
- 28/06/2017 ■ **MOCK TRIAL AND ROUNDTABLE SESSION** " Ethique, algorithmes and justice "
> **Law Department of Université Catholique de Lyon and Lyon's Administrative Court of Appeal**
- 28/06/2017 ■ **STUDY DAY** "Admission Post-bac, textbook case of public algorithms"
> **Fondation Internet Nouvelle Génération (FING) and Etalab**
- 03/07/2017 ■ **DAY** "Algorithms and digital sovereignty"
> **Allistene's CERNA**
- 05/07/2017 ■ **DAY** "Ethics and artificial intelligence"
> **French National Center for Scientific Research (CNRS)'s ethics committee (COMETS) and French Association for AI (AFIA)**
- 22/08/2017 ■ **DEBATES** on algorithms in education.
- 24/08/2017 ■ > **Ligue de l'Enseignement (Education League)**
- 05/09/2017 ■ **DEBATE MORNING** "Work in the algorithm era: what ethics for employment?"
> **Renaissance Numérique and Randstad**
- 11/09/2017 ■ **SYMPOSIUM** ""Convergences of law and digital technology"
> **Bordeaux University**
- 13/09/2017 ■
- 14/09/2017 ■ **DAY** "Algorithms and Politics. The ethical issues of forms of digital computing from the perspective of social sciences"
> **School of Advanced Studies in the Social Sciences (EHESS) and Complex Systems Institute of Ile-de-France**

- 15/09/2017 ■ **DAY** on healthcare research into its ethical and regulatory aspects (data, algorithms)
> **Necker Hospital and Institut Imagine**
- 15/09/2017 ■ **ROUNDTABLE SESSIONS** "Algorithms and risks of discrimination in the insurance sector"
> **Ligue des Droits de l'Homme (Human Rights League)**
- 20/09/2017 ■ **SYMPOSIUM** "Ethical issues raised by algorithms"
> **INNOvation Ouverte par Ordinateur (INNOOO)**
- 20/09/2017 ■ **DEBATE MORNING** "Are the ethics of algorithms and AI compatible with value creation in the IoT?: Internet of Things and/or Internet of Trust?"
> **ESCP Europe (IoT Chair)**
- 20/09/2017 ■ **SYMPOSIUM** "Ethics and digital technology"
> **Collège des Bernardins**
- 21/09/2017 ■ **DEBATE** "Opportunities and challenges of advanced machine learning algorithms"
> **The John F. Kennedy Jr. Forum at Harvard Kennedy School**
- 21/09/2017 ■ **SYMPOSIUM** "Lex Robotica (on the boundary between robotics and Law: Designing the humanoid in 2017)"
> **Conservatoire National des Arts et Métiers (CNAM)**
- 22/09/2017 ■ **ROUNDTABLE SESSION** "AI and ethics of algorithms"
> **Fotonower**
- 26/09/2017 ■ **SYMPOSIUM** "Predictive algorithms: what are the ethical and legal issues?"
> **Research Centre of the National Gendarmerie School of Officers (CREOGN)**
- 28/09/2017 ■ **CONSULTATION** "What future for medicine in the age of artificial intelligence?"
> **Rhône Département-level Council of the Medical Association**
- 29/09/2017 ■ **ROUNDTABLE SESSIONS** "Ethics of algorithms and big data"
> **French Development Agency (AFD) and Caisse des dépôts et consignations (CDC)**
- 04/10/2017 ■ **SYMPOSIUM** "Algorithms and the battlefield"
Debate-forum "Towards a friendly form of Artificial Intelligence?"
> **National Defence Institute of Higher Studies (IHEDN)**
- 06/10/2017 ■ **DEBATE-FORUM** "Towards a friendly form of Artificial Intelligence?"
> **Laboratory for Collective and Artificial Intelligence (LICA)**
- 12/10/2017 ■ **ROUNDTABLE SESSION** "Law and artificial intelligence: what responsibility(ies)?"
> **Club des Juristes and Microsoft**
- 14/10/2017 ■ **PUBLIC CONSULTATION** on the ethical issues raised by algorithms
> **CNIL**
- 07/09/2017 ■ **PUBLIC CONSULTATION** on the governance of artificial intelligence
31/03/2018 > **The Future Society at Harvard Kennedy School**

GLOSSARY

Algorithm

Description of a finite and unambiguous sequence of steps or instructions for producing a result (output) from initial data (input).

Artificial intelligence (AI)

Theories and techniques involving «making machines do things that would require intelligence if done by men” (Marvin Minsky). Weak AI (AI capable of simulating human intelligence for one specific task) is distinguished from strong AI (autonomous, artificial general intelligence which could apply its capacities to any problem, in this way replicating a strong characteristic of human intelligence – i.e. of a form of machine “consciousness”).

Big data

Refers to the conjunction between, on the one hand, huge volumes of data that have become difficult to process in this digital age and, on the other, the new techniques which are enabling such data to be processed – and even unexpected information to be inferred from it by identifying correlations.

Chatbot

A computer program which converses with its user (for example, empathetic robots to assist patients, or automated conversation services in customer relations).

Machine learning

Current application of artificial intelligence, based on automated methods whereby computers can acquire and learn new knowledge, and therefore operate without being explicitly programmed.

Supervised machine learning

The algorithm learns from input data labelled by humans and then defines the rules based on examples which are validated cases.

Unsupervised machine learning

The algorithm learns from unlabelled input data and carries out its own classification; it is free to produce its own output when presented with a pattern or variable. A practice which requires trainers to teach the machine how to learn.



Commission Nationale de l'Informatique et des Libertés

3 place de Fontenoy
TSA 80715
75334 PARIS CEDEX 07

Tél. 01 53 73 22 22
Fax 01 53 73 22 00

www.cnil.fr

