Research Ethics in Machine Learning

Cerna Collectif

Research Ethics in Machine Learning

CERNA Report, february 2018

Research Ethics Board of Allistene,
the Digital Sciences and Technologies Alliance

Allistene, the Digital Sciences and Technologies Alliance, was founded in 2010 by major national players in ICST research: the CDEFI (Conference of Engineering College and Training Directors), the CEA (French Alternative Energies and Atomic Energy Commission), the CNRS (National Centre for Scientific Research), the CPU (Conference of University Chairpersons), the IMT (Institut -Mines-Télécom, group of engineering and management graduate schools) and Inria (the National Institute for computer science and applied mathematics).

Allistene's aim is to promote collaborative thinking and coordinate its members in terms of foresight and overall strategy, the challenges of advanced training, planning, European and international partnerships, and industrial utilization and liaison. Its functioning is based on working groups called «programmatic groups», comprising top-flight researchers.

www.allistene.fr

cdefi Conférence des Directeurs des Écoles Françaises d'Ingénieurs

cea

cnrs

cpu Conférence des Présidents d'Université

Inria

Institut Mines-Télécom

RESEARCH ETHICS IN MACHINE LEARNING

PREFACE

The rapid spread of innovation-based IT practices complicates the interaction between technological capacity and societal adoption and reduces the relevance of forecast activities about the consequences of research. However, this relative unpredictability does not free scientists of responsibility, but should instead motivate ethical reflection and the quest for appropriate perspectives and methods. Researchers should be aware that their work de facto contributes to changing society and humanity, and the process is not always predictable. Although the responsibility for this impact should not be borne by them alone, they too have a share of collective responsibility. Against this background, the aim of CERNA is to encourage and support researchers in the exercise of ethical reflection about their work.

This document is addressed to IT researchers, developers, and designers. Societal issues are listed but not explored in depth. CERNA considers only scientifically plausible possibilities, avoiding science-fiction scenarios that might become a source of confusion.

MEMBERS OF THE MACHINE LEARNING WORKING GROUP

Laurence Devillers,
Professor at Paris-Sorbonne 4, LIMSI-CNRS, CERNA, leader of the team

Serge Abiteboul,
Research Director at Inria, ENS-Paris, Member of the Academy of Sciences

Danièle Bourcier,
Emeritus Director of Research at CNRS, CERSA, CERNA

Nozha Boujemaa,
Research Director at Inria

Raja Chatila,
Professor at UPMC, Director of ISIR, CERNA

Gilles Dowek,
Research Director at Inria, ENS-Saclay, CERNA

Max Dauchet,
Emeritus Professor, University of Lille, Chairman of CERNA

Alexei Grinbaum,
Researcher at CEA, IRFU/LARSIM, CERNA

With the collaboration of Christophe Lazaro, researcher at University of Namur, CERNA, and of Jean-Gabriel Ganascia, Professor at UPMC Paris 6, LIP6, CERNA

PEOPLE CONSULTED (SUB-GROUP)

Alexandre Allauzen,
Lecturer, Université Paris 11, LIMSI

Edouard Geoffrois,
Head of Program at the Department of Information and Communication Sciences and Technologies, ANR

Mathieu Lagrange,
Researcher at CNRS, LS2N

Arnaud Lallouet,
Chief Engineer, Huawei Technologies Ltd.

Olivier Teytaud,
Researcher, Inria

PEOPLE CONSULTED AT CERNA'S LEARNING AND AI DAY, JUNE 13, 2016, PARIS[1]

Tristan Cazenave,
Professor, Paris-Dauphine, LAMSADE

Milad Doueihi,
Chair of Digital Humanism, Paris-Sorbonne; current holder of the Chaire des Bernardins on *The Digital Challenge to the Human*

Benoît Girard,
Research Director at CNRS, ISIR

Jean-Baptiste Mouret,
Researcher at Inria, Larsen Team

EXPERT PROOFREADER

Léon Bottou,
Facebook AI Researcher: machine learning, artificial intelligence

[1]http://cerna-ethics-allistene.org/journ%C3%A9e+apprentissage/.

INTRODUCTION

Automatic learning, also called statistical learning and commonly known as machine learning, has recently made spectacular advances, headlined in 2016 by the victory of the AlphaGo program over the world Go playing champion, Lee Sedol. Machine learning has multiple applications—e.g. search engines, image and speech recognition, automatic translation, chatbots—which are beginning to appear in sectors such as health, energy, transport, education, commerce and banking.

The successes of machine learning, one of the fields of artificial intelligence (AI) research, arise out of increases in computing power and data storage and processing capacity ("big data"). They have been followed by sensationalist and inaccurate media stories suggesting that machines—sometimes robots—could replace human beings. While this scenario remains beyond the reach of today's science, it is nevertheless true that there needs to be ethical attention to the proper use of learning algorithms and increasingly complex, large, and ubiquitous volumes of data. Initiatives along these lines, both public and private, at national, European or international levels, have been emerging since 2015.

Against this background, the purpose of the present document is to:

- Raise awareness and provide "researchers" with food for thought and certain waymarks. *For reasons of convenience, the term "researcher" is used here to refer to people—designers, engineers, developers, entrepreneurs—and their communities or institutions;*
- Contribute to a wider debate on the ethical and societal questions associated with the development of artificial intelligence,

so that machine learning develops to the benefit of society.

CERNA is therefore addressing two kinds of reader here: on the one hand specialists, and on the other hand anyone interested, whether decision-makers or ordinary citizens.

Part I introduces the core concepts of machine learning and illustrates them through the specific method of multi-layer neural networks and

deep learning. Part II describes examples of use that are already widespread or destined to become so. These two parts provide a technological substrate for the ethical reflections and are intended in particular for non-specialists. Part III presents the general ethical questions associated with digital systems and identifies those specifically linked with machine learning.

Part IV analyses these ethical questions and makes recommendations addressed to the scientists and communities that design and develop machine learning systems. These recommendations draw attention to points where individual and collective ethical attention is called for, but they should in no way be seen as "recipes". They are articulated around six questions:

1. What data are selected/used for the machine to learn from?
2. Can we be sure that the machine will only perform the tasks for which it was designed?
3. How can we assess a system that learns?
4. What decisions can and cannot be delegated to a machine learning system?
5. What information should be given to users on the capacities of machine learning systems?
6. Who is responsible if the machine malfunctions: the designer, the owner of the data, the owner of the system, its user, or perhaps the system itself?

The initiatives described in Part V illustrate the topicality of the ethical questions associated with developments in machine learning and more generally in artificial intelligence. Part VI concludes with general recommendations addressed to people in the scientific community and society's decision-makers.

I. WHAT IS MACHINE LEARNING?

One of the goals of artificial intelligence researchers is to construct systems with the ability of perception, learning, abstraction, and reasoning. To achieve this, learning algorithms use different statistical methods based on training data, for example in order to construct rules of deduction and decision trees, or to configure neural networks, and then apply them to new data.

To predict a phenomenon from past observations presupposes a causal mechanism. Explaining that mechanism is not always easy. Machine learning is a statistical approach that can discover significant correlations in large masses of data, in order to build a predictive model when it would be difficult to construct an explanatory model. Handwriting recognition is an example of a problem that is difficult for a machine. In order to recognize a letter or a number, some algorithms use preset rules, but others "learn" to recognize the letters of the alphabet from a large number of examples. These algorithms, which use data to learn to solve a problem, are called "machine learning algorithms". They are being developed for application in many fields, such as finance, transportation, health, well-being, even art.

In transportation, for example, systems obtained by machine learning are used to enable autonomous vehicles to visually recognize their environment. In a quite different field, the face recognition made popular by GoogleFace and Facebook are used in social networks to identify people in photos. In the world of games, IBM's *Deep Blue* beat the world chess champion back in 1997.[2] In 2011, IBM's *Watson took* part in three rounds of the American TV quiz *Jeopardy*, ultimately winning the game.[3] In 2016, *Alphago of Google DeepMind* defeated one of the world's top Go players, Lee Sedol.[4]

There is now an emerging field of research dedicated to improving the explainability and transparency of machine learning systems, together with their contextual adaptation and the match between what they learn and what human beings expect of them. The goal of this research is thus to go beyond the simple use of machine learning to build models without understanding, and to try to explain those models.

[2]Deep Blue was a specialist chess playing supercomputer developed by IBM in the early 1990s.
[3]Watson is an artificial intelligence computer program designed by IBM for the purpose of answering questions formulated in natural language.
[4]AlphaGo is a computer program designed to play the game Go, developed by the British company Google DeepMind.

I.1 The different types of machine learning algorithms

Machine learning algorithms are multiple and diverse. They can be divided into three main categories, depending on whether their learning method is *supervised, unsupervised*, or based on *reinforcement*.

In *supervised learning*, the training data (the data used by the machine to learn) must first be labeled by "experts". For example, in order to build a system for recognizing letters in images, the experts label images for all the data that represent the letter "a", "b", etc. In the initial, so-called learning phase, the machine constructs a "model" of the labeled data, which can be a set of rules, a decision tree, a set of matrices in the case of neural networks, etc. This model is then used in the second, so-called recognition phase, where for example the algorithm recognizes a letter in a new image. Support vector machines (SVM),[5] or neural networks such as multi-layer perceptron systems using backpropagation learning with gradient descent,[6] are examples of supervised machine learning.

In *unsupervised learning*, there is no need for an expert to label the data. The algorithm discovers the structure of the data on its own, by classifying them into homogeneous groups. *K-means clustering* (a method of partitioning data)[7] and neural networks such as Kohonen maps (a method of reducing dimensionality)[8] are examples of unsupervised machine learning algorithms.

In *reinforcement learning*, the goal is for the machine to learn from experience what needs to be done in different situations, in order to optimize a quantitative reward over time. The algorithm works by trial and error, with each error prompting it to improve its performance in solving the problem. Here, the role of the expert is limited to setting the success criteria for the algorithm. TD-learning[9] and Q-learning[10] are examples of reinforcement learning algorithms.

[5]Boser Bernhard E., Guyon, Isabelle M., Vapnik, Vladimir N., "A training algorithm for optimal margin classifiers" COLT'92, pp.144-152
[6]Rumelhart, David E. Hinton, Geoffrey E., Williams, Ronald J. (8 October 1986). «Learning representations by backpropagating errors». Nature. 323 (6088) : 533–536
[7]MacQueen, J. B. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281– 297 (1967)
[8]Kohonen, Teuvo (1982). «Self-Organized Formation of Topologically Correct Feature Maps». Biological Cybernetics. 43 (1) : 59-69
[9]Sutton, R.S., 1988, Learning to Predict by the Method of Temporal Differences, Machine Learning, 3, pp. 9-44
[10]Watkins, C.J.C.H. & Dayan, P., Q-learning, Mach Learn (1992) 8 : 279

Finally, there are also intermediate methods, such as semi-supervised learning, which sometimes leaves room for human intervention, but raises real-time constraints. Moreover, several learning methods are often combined within a single system.

I.2 An example: multi-layer neural networks

Multi-layer neural networks are trained using machine learning algorithms. In very broad terms, their design was originally inspired by biological neurons, using a concept of artificial neurons based on the analogy with neurons in the brain. Like a natural neuron, an artificial neuron (Figure 1 below) has multiple input values that determine a single output value, which is then propagated as an input to other neurons.

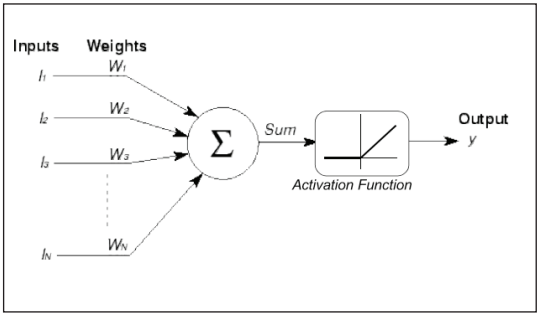

Figure 1 : The McCulloch-Pitts artificial neuron is a very simple mathematical model derived from an analysis of biological neuron function.

This output can be a simple linear combination of the inputs

$$y = w_1\ input_1 + w_2\ input_2 + \ldots + w_n\ input_n$$

or a composite of this kind of linear combination and an activation function (threshold function, sigmoid function, etc.). The synaptic weights w_1, w_2... of each neuron are determined iteratively during the phase of learning on labeled data. The capacity for these weights to change in neurons over time is called "synaptic plasticity".[11]

The first neural network, Rosenblatt's Perceptron, dating back to the 1950s, had only one layer. Some systems such as the multi-layer perceptron

[11] Hebb, D.O., The Organization of Behavior, New-York, Wiley and Sons, 1949

consist of several layers of artificial neurons, which enable them to recognize a shape such as a picture, but not to infer human concepts or the logic connecting them. A layer may consist of thousands of neurons, and therefore millions of parameters. Between the input layer and the output layer, the network may contain several dozen so-called hidden layers. "Deep learning" systems are neural networks that contain a large number of layers.[12]

The learning phase determines the values of the synaptic weights from a very large sample of data (possibly millions of items). In the supervised *gradient descent backpropagation learning algorithm*,[13] the difference between the expected outputs and the actual outputs is reduced step-by-step *(gradient descent)* by changing the parameters from the output upto the first layers *(backpropagation)*, until a local minimum is obtained (the absolute minimum is difficult to achieve). The initial value of the synaptic weights is sometimes randomly drawn, but an unsupervised learning algorithm can also determine it.

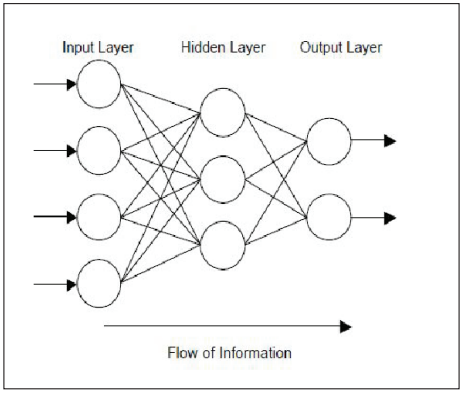

Figure 2 : The multi-layer perceptron, here with a hidden layer.

[12]LeCun, Y. Bengio and G. Hinton, (2015) "Deep learning", Nature, vol. 521, no 7553, 2015, p. 436–444 (PMID 26017442, DOI 10.1038/nature14539
[13]Rumelhart, David E., Hinton, Geoffrey E., Williams, Ronald J. (8 October 1986). "Learning representations by backpropagating errors". Nature, vol. 323 n° 6088, p. 533–536

Designing a deep network that can learn to perform satisfactory classification—where the term "satisfactory" is to be understood empirically, in the sense that the results obtained with real-world data meet expectations—requires a great deal of expertize and engineering. As Yann Le Cun explains, deep learning exploits the modular structure of real-world data.[14] Its success lies in its capacity to learn without the need for an explicit data model. The mathematical framework of gradient descent explains these methods, but does not guarantee successful learning (convergence theorems only exist in very simple cases) and these algorithms require a large number of iterations in order to converge empirically on an acceptable solution. The recent success of these methods owes much to the increase in computing power and to the large volumes of data available.

The architecture (neuron types, connection choices) must be adjusted to the field of application. Deep learning systems can have different architectures, for example recurrent or convolutional networks, and complex approaches that combine several deep learning systems.

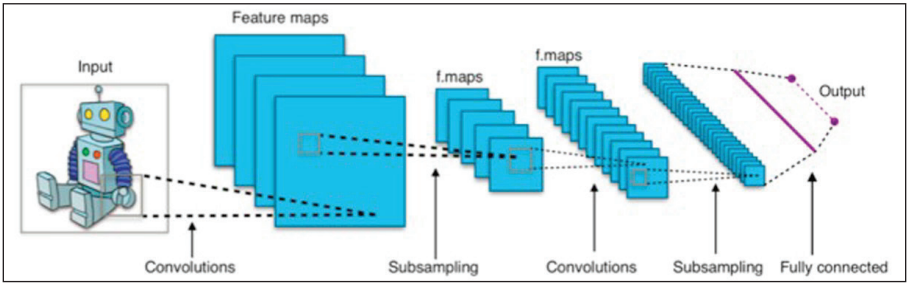

Figure 3 : Standard architecture of a convolutional network[15]

For example, the AlphaGo system developed by Google DeepMind combines Monte Carlo tree search and Deep Learning. In this application, a deep network is trained on games played by top human players in order to learn to predict their moves. This network reaches 3rd Dan level and improves its game through reinforcement learning, playing 30 million

[14]Yann Le Cun, Chair in Computer and Digital Sciences at Collège de France/Inria, 2015-2016, https://www.college-de-france.fr/site/yann-lecun/

games against itself. It then uses the results of these games to train another network, which learns to assess positions. Monte Carlo tree search uses the first network to select interesting moves and the second network to assess positions. Serge Abiteboul and Tristan Cazenave describe the typology of AlphaGo's deep networks and the Monte-Carlo principle in the 2016 SIF report:[16] *"The networks used for AlphaGo, for example, are made up of 13 layers and between 128 and 256 feature maps. For specialists: they are "convolutional", with 3x3 filters, and use Torch, a language derived from the Lua programming language (…) The principle of Monte-Carlo search is to collect statistics on possible moves from randomly played games. In fact, the games are not completely random, and choose moves with probabilities that depend on a shape, on the context of the move* (arrangement of the stones on the goban)*. All the states encountered in the random games are memorized and the statistics on the moves played during those states are also memorized. This means that when the algorithm returns to a previously encountered state, it chooses the moves with the best statistics. AlphaGo combines deep learning with Monte-Carlo search in two ways. Firstly, it uses the first network, which plans the moves, to try those moves initially in random games. Then, it uses the second network, which assesses the positions, to correct the statistics derived from the random games."*

The deep learning field accounts for a significant proportion of articles published in the main machine learning journals. Yann Le Cun gave a course on this topic within the framework of the "Collège de France"s annual chair in "Computer and Digital Sciences".

Artificial Intelligence Platforms[17] make available a wide range of networks that non-specialists can use to test or develop machine learning applications. Other initiatives exist that seek to bring learning platforms within everyone's reach. Google provides free access to its *DeepDream* Deep Learning testing platform, as well as to its open source machine learning tool, *TensorFlow*. Facebook is doing the same with its open source machine learning hardware, Big Sur, which can run large neural networks. In addition, *OpenAI* was recently set up with significant private funding (US$1 billion), notably provided by Elon Musk, Peter Thiel, and Reid Hoffman, who state that "Our goal is to advance digital intelligence

[15] "Convolutional Neural Networks (LeNet) — DeepLearning 0.1 documentation" on DeepLearning 0.1, LISA Lab
[16] "Go : Une belle victoire... des informaticiens !", Serge Abiteboul, Tristan Cazenave. Bulletin n° 8 de la SIF 2016
[17] https://www.predictiveanalyticstoday.com/artificial-intelligence-platforms/

in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return." Initiatives of this kind encourage acculturation and the collaborative development of learning tools, and the purpose of the open source approach is to ensure collective control. However, they raise the possibilities of a proliferation of poorly controlled and insecure applications developed by some individuals and start-ups.

II. EXAMPLES OF MACHINE LEARNING APPLICATIONS

"It is very likely that by the time you read these lines, you will already have used the results of machine learning algorithms several times today: your favorite social media network has perhaps suggested new friends, and your search engine has decided that certain web pages are relevant to you but not to your neighbor. You have dictated a message on your mobile phone, used an optical character recognition program, read an article that was specifically suggested to you on the basis of your preferences and which may have been automatically translated." (Colin de la Higuera, *Binaire* blog in the newspaper *Le Monde*, June 23, 2015)

In fact, many artificial agents use machine learning modules. These agents may be software entities like chatbots, or hardware entities like robots or autonomous cars. They can vary in their degree of autonomy and some may appear to be social actors, with the capacity to interact, and even to simulate emotions or make decisions.

II.1 Personalized recommendations

The traces we leave through our web searches and through the objects we are connected to are used by learning algorithms to identify our shopping preferences, our lifestyles, and our opinions. By contrast with algorithms that simply collect statistics, these have—or can have—the capacity to make individual recommendations. So when we browse the Internet and buy things online, we don't realize that our digital trail can prompt algorithms to categorize us in ways that may affect our insurance premiums or trigger lifestyle recommendations. This means that compliance, transparency, trust, and fairness are crucial properties of the learning algorithms underlying these processes.

II.2 Chatbots

Chatbots, or conversational agents, are software agents that can automatically process natural language conversation. They are increasingly used as personal assistants or for handling e-commerce transactions running on IT platforms. They may even be responsible for the majority of online "chats" with human beings. The mass proliferation of these interactions, with no hierarchy or clear distinction between human and machine, could ultimately influence the corpus of texts available online. Moreover, chatbot behavior is conditioned by training data. The UK's National Health Service has been running experiments with learning bots since the beginning of 2017, not only to reduce call congestion, but also in the hope that linking these bots to very large medical databases will improve the health advice service it provides. However, bots can also be trained or used for nefarious purposes. They are already employed to exercise influence, both in the commercial sphere and in electoral politics. In April 2016, Microsoft's Tay chatbot, which had the capacity to learn continuously from its interactions with web users, started to spout racist language after just 24 hours online.[18] Microsoft quickly withdrew Tay.

The two examples that follow, covered in CERNA's position paper on Research Ethics in Robotics (2014),[19] are outlined briefly here in connection with machine learning.

II.3 Autonomous vehicles

Any accident involving a totally or partially autonomous vehicle triggers a massive media response.[20] Yet of the USA's 10 million annual road traffic accidents, 9.5 million are caused by human error, and it is likely that traffic flows consisting of autonomous cars would be safer than those of cars driven by people. At present, all autonomous cars of the same type are delivered with the same configuration, and stop learning once they are on the road, but it is a safe prediction that their successors will retain the capacity to learn continuously from their environment. This means that it will be crucial for their behavior to be regularly assessed.

[18] http://www.lemonde.fr/pixels/article/2017/04/15/quand-l-intelligence-artificielle-reproduit-le-sexisme-et-le-racisme-des-humains_5111646_4408996.html

[19] https://hal.inria.fr/ALLISTENE-CERNA/hal-01086579v1, 2014

[20] E.g. http://www.lexpress.fr/actualite/monde/amerique-nord/premier-accident-mortel-pour-une-voiture-tesla-en-pilote-automatique 1808054.html and http://www.lefigaro.fr/secteur/high-tech/2016/03/01/32001-20160301ARTFIG00118-la-google-car-provoque-son-premier-accident-de-la-route.php

II.4 Robots that interact with people and groups

The ability to adapt to the environment that comes with the capacity to learn should, in the future, foster the use of robots that works with people, in particular as companions or carers. The construction of "social" robots to provide personal care will require their use to be controlled, especially when they are interacting with sick or elderly people.

III. ETHICAL ISSUES

Traditional ethical theories are instantiated in new forms in digital technology and machine learning. The dilemmas associated with autonomous cars are an example that has prompted extensive commentary.[21] To put it simplistically, an autonomous vehicle that had to choose between sacrificing its young passenger, or two incautious children, or one elderly cyclist going about his lawful business, could be programmed to apply Aristotle's virtue ethics—in this case abnegation—by sacrificing its passenger, deontological ethics entailing compliance with the highway code by sacrificing the children, and consequentialist ethics if it sacrifices the cyclist—in this case by minimizing the number of years of life lost.

The purpose here is not to tackle such issues, which are questions for society as a whole, but to direct the researcher's attention, in the case of machine learning, to certain specific properties that the behavior of a digital system must fulfil.

For any digital system, the aim should be to embody the properties described in III.1. However, machine learning systems possess certain specificities, described in III.2, which come into conflict with those general properties.

III.1 General properties of digital systems

- **Trustworthiness and fairness:** when applied to computer systems, trustworthiness means that those systems, when in operation, behave as their designers claim. If, for example, the designers claim that a system does not store its users' personal data, it must not do so. A computer system is fair if it treats all its users equitably.

[21] Jean-François Bonnefon, Azim Shariff2, Iyad Rahwan, The social dilemma of autonomous vehicles, Science 24 Jun 2016

• **Transparency, traceability, and explainability:** a system is transparent if its operation is not hidden, e.g. if it is possible for a user to monitor its behavior. This transparency depends in particular on traceability, the availability of sufficiently detailed information on its actions (stored in a log) for those actions to be subsequently backtracked. Traceability is essential firstly for the attribution of responsibility, i.e. potentially as the basis for legal proceedings, and secondly for the diagnosis and correction of dysfunctions. Traceability also ensures that a system's operation can be explained from the tracks it leaves, hence the quality of explainability.

• **Responsibility:** in order to be able to attribute liability in the event of a dysfunction, it must be possible to distinguish two agents: the system's designer, and it user. The originator or the designer is responsible if the system is poorly designed, the user is responsible if he or she has misused the system (just as, with the use of a hammer, the user is responsible if he clumsily hits his fingers, whereas the designer is responsible if the head flies off and knocks the user out), with the proviso that it is the professional's duty to provide any nonprofessional (the user) with additional information.

• **Compliance:** A digital system must fulfil its specifications, and its specifications must be in compliance with the law. A system's compliance with its specifications means that it is designed to carry out specified tasks within the constraints set out in those specifications. The requirements of the specifications often constitute a restrictive interpretation of the law, since they are unable to translate legal nuances. Compliance must be verified before the system is used, through an analysis of its code and data. This means, for example, compliance with data protection rules in a data analysis system, or compliance with the Highway Code in an autonomous car.

III.2 Some particular features of machine learning systems

• **Specification problem:** The purpose of machine learning is precisely to tackle tasks that cannot be specified in formal computing terms. Machine learning replaces such formal specification with a model whose parameters the machine sets empirically from a mass of data. For instance, if we want to design a program to recommend books to readers in a public library, there are two possible approaches. The first uses explicit rules: this kind of program could, for example, define three lists

of books, aimed respectively at children, teenagers and adults, ask the reader her age and, depending on whether the answer is under 12, between 13 and 17, or over 18, randomly pick a book from one of those lists. In this case, the program is easy to specify: the recommended book simply needs to be in the category that corresponds to the reader's age. The second approach uses a learning algorithm that works differently, starting with two parameters: the reader's age and a list of books that readers of her age say they have liked. In this case, the list of books recommended for each of the age categories is dynamic, varying during the training of the algorithm. An advantage of this is that, instead of relying on a rough division of the population into three broad categories, the recommendations can be much more targeted. A disadvantage is that a malicious trainer could train the algorithm to recommend books that are unsuitable for the reader's age. In this case, in the absence of any advance categorization, the specification "The recommended book must belong to the category that corresponds to the reader's age"—would be meaningless. Moreover, in this example, we assume that the learning algorithm not only dynamically constructs the list of books to recommend to readers on the basis of their age, but also constructs the "categories" used to establish those recommendations. So it might not employ the usual concepts (age, gender, etc.), but concepts specific to itself, which may not necessarily mean anything to the human trainer, making it even harder to specify what is expected of the algorithm. For example, the category "readers who borrow a book between 3 pm and 3:15 pm" might be relevant to the machine, but would seem arbitrary or meaningless to a human user. In the case of a public library, it is essential that the recommendations should be explainable. The categories that lead to an outcome, even if they emerge from a learning process, must be expressed in human language and clearly specified.

• **Training agent:** Apart from the designer and the user, machine learning systems introduce a third type of agent, which uses a dataset to train the machine learning system. So as well as being caused by bad system design or use, in which case the designer or user should be held responsible, a dysfunction could also arise from poor training, in which case it is the trainer that should be held accountable. This situation is not entirely new. There are also three agents involved when the writer of a program uses a compiler: the designer of the compiler, equivalent to the designer of the learning system; the program's author, equivalent to the trainer; and the program user, equivalent to the user of the machine

learning system. In this situation, the program both transforms and is transformed: it transforms the inputs, but is itself transformed by the compiler. However, with machine learning the reality is even more complex. In a system that learns continuously, all the users are also trainers. Upstream from or with the agreement of the trainer, data can be acquired from sensitive questions or be subject to processing restrictions or exclusions (personal data, image rights, etc).

• **Learning without understanding:** Automatic learning algorithms can beat the best players at chess or Go, but they are often incapable of explaining why they played one move rather than another, because this "explanation" is based on the adjustment of millions of synaptic weights and not on simple concepts that humans can understand.[22] Similarly, one of the strengths of learning-based image recognition algorithms is that they can recognize a chair without necessarily employing concepts such as a chair leg, seat, or back... but it also means that it is hard for the algorithm to explain why it identified a chair in an image. The problem is even more complex in the case of an unsupervised algorithm, which learns without reference to any goal that humans can understand, or of a reinforcement algorithm, which seeks to optimize a reward function that is often too simple to provide an explanation of how the stated goal is achieved. The correlations between the concepts learned (clusters or indexing vocabulary) and the zones of an analyzed image sometimes differ widely between human and machine: the regions of an image to which networks and human beings pay attention when answering a question are not the same[23].

• **Dynamically evolving models:** When the system continues to learn after deployment, its long-term behavior is difficult to control. During use, the system may learn behaviors that introduce bias, thereby breaking the criterion of fairness. For example, a personal robot may start to behave abusively in its interactions with humans, or a lending algorithm may begin to discriminate against minorities or particular social groups in its offers. Moreover, the algorithm itself can generate unforeseen categories that can prove discriminatory with respect to fundamental freedoms (e.g. the use of non-significant risk selection criteria for credit scoring, such as the applicant's height). It is not always easy to track how the data used for learning are collected.

[22] A human Go player may also be unable to explain a move. In the case of games, an absence of explainability is not a significnt issue.

[23] Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, Batra, 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, arXiv :1606.03556

• **Learning instability:** deep learning is currently one of the most effective approaches in the field, but the algorithm can nevertheless present some instability. A change—even imperceptible to the eye—in a small number of pixels in a picture that a human being can identify, for example a photo of a car, can make that picture unidentifiable to a deep learning system.[24] Conversely, some images that mean nothing to human beings are may be labeled automatically as close to learned shapes. So images that have very different meanings for human beings can be indexed as the same: a classic example is the photo of a panda identified as a gibbon.[25] The outputs in a deep neural network assign a confidence value to a recognition; a gradient method is then used to increase that confidence value, keeping the network parameters the same but making step-by-step changes to the input, thereby converging on an input that produces the output in question with a maximum confidence value. Using this method, Google's Deep Dream platform can easily alter photos to make them look surreal.

• **Assessment and control:** Since it is difficult to formulate the specifications for a system that learns, such a system is difficult to assess. On the other hand, its effects can be assessed retrospectively. For example, it is difficult to judge whether an autonomous vehicle accelerates or brakes at the right time, but it is possible to gage retrospectively whether the vehicle has caused fewer or more accidents than a human driven vehicle. When machine learning systems continue to evolve while in operation, they need to be assessed at regular intervals throughout the period of use. Different types of agents could appear in the management of learning systems: "interpreter" agents that use sets of tests to help understand the machine's behavior, agents that "evaluate" or "check" the learning algorithms in order to ensure that the systems remain trustworthy and fair, "legal" agents which ensure that the systems operate within the law.

These specificities, which are all points on which researchers need to be vigilant, are the counterpart to the range of possibilities opened up by machine learning. They open the way to new research that will lead, in numerous sectors, to the development of systems that will be more reliable than human beings, who are also learning entities that fail and make mistakes.

[24]Intriguing properties of neural networks, Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus https://arxiv.org/abs/1312.6199
[25]Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, Explaining and harnessing adversarial examples, ICLR 2015 https://arxiv.org/pdf/1412.6572v3.pdf

IV. RECOMMENDATIONS FOR LEARNING SYSTEMS IN SIX THEMES

IV.1 Learning system data

It is data that governs the outcomes of learning algorithms. Because this data is captured in real-world conditions, and because there is no preliminary model, it is difficult to assess whether it meets the desired objectives, and data bias—whether intentional or not—can have serious consequences **[DAT-1]**. Training data can be discriminatory, for example favoring certain physical features in a face recognition application, or may reflect political, religious, or other preferences of system designers or trainers. In 2015, for example, Google put a face recognition system online that worked better with pale skins, obliging the company to make a public apology.[26] In certain cases, biases can be illegal, for example offering less beneficial financial products to members of minority groups. As an example of commercial bias, the European Union recently sanctioned Google for skewing the results of its search engine in favor of its own services.[27]

Fairness is difficult to specify. It may be based on subjective or conflictual criteria. Different cultures may, for instance, prioritize different criteria, for example favoring either equality or reward for merit (the notion of "equality of opportunity" illustrates the complexity). At the individual level, if fairness means as far as possible giving people what they want, should a robot fulfil an elderly person's request for several whiskeys, even if this is not good for their health? **[DAT-2]**.

The law may prohibit the use of particular variables, such as ethnicity, sex, age, or religion, for deciding to provide or withhold certain services for specific people. Nonetheless, an algorithm might reconstruct the values of such variables, and then take decisions based on them **[DAT-3]**.

Traceability ensures the possibility of "tracking" data captured from the environment or exchanged in the course of certain events, as well as the computations carried out by the system. It is essential to transparency and to the analysis of functional or dysfunctional performance. If the system's code and data are open—two key factors of transparency—verification is obviously facilitated (though still difficult to carry out). In the case of learning machines, one should make efforts to keep records

[26]http://www.huffingtonpost.fr/2015/07/02/logciel-reconnaissance-faciale-google-confond-afroamericains-gorilles_n_7711592.html
[27]http://europa.eu/rapid/press-release_IP-16-2532_fr.htm

of training data, and of the conditions of its collection and validation. If the system continues to learn once in operation, tracking is more complicated. It is essential that the traces be monitored in order to detect deviations from expected behaviors. In such cases, the system can ask a human being to check the learning status of the system. Researchers should remember that learning traces are data and for this reason need to comply with data privacy rules, even if they are to be used exclusively for technical monitoring purposes **[DAT-4]**.

Priorities and recommendations

[DAT-1] Quality of training data

The designer and the trainer should pay attention to the training data and the conditions of data capture throughout the operation of the system. Trainers of machine learning systems are responsible for the presence or absence of bias in the data used in learning, in particular, for "continuous" learning, i.e. that takes place while the system is in use. In order to check the absence of bias, they must rely on measurement tools that have yet to be developed.

[DAT-2] Data as a mirror of diversity

When selecting data, trainers of machine learning systems must ensure that those data reflect the diversity of the groups of users of those systems.

[DAT-3] Variables in which the data pose a risk of discrimination

The trainers (who may also be the designers or users) should pay attention to protected variables, e.g., variables that may permit social discrimination. These variables, such as ethnicity, sex or age, must not be used or be regenerated based on correlations. Personal data must also be protected as required by existing legislation.

[DAT-4] Tracking

Researchers must ensure that machine learning is traceable, and provide protocols for that purpose. The traces are themselves data, and as such also demand ethical handling.

IV.2 Autonomy of machine learning systems

For a digital system, autonomy "is the capacity to operate independently from a human operator or from another machine by exhibiting nontrivial behavior in a complex and changing environment. (…)"[28] Autonomy is a relative concept. The autonomy that a system can achieve depends firstly on the complexity of the environment, which can be measured by the quantity and variety of the information it contains and of the flows and dynamics of that information, and secondly on the complexity of the task, which depends on the structure of all the system's possible states (state-space). If the environment of use of an autonomous system is complex, such as a city street in the case of an autonomous car, preliminary learning is often needed. If the environments of use are changeable or unpredictable, as in the case of a companion robot, personalized learning is required, which may need to be updated periodically or to continue during the entire period of use.

For a machine endowed with autonomy to operate in a way that is faithful to the intentions of its designers and operators, the machine's internal representation of a situation and the behavior it manifests in response must be intelligible and in keeping with what its operator or human user expect. Recommendations on this are formulated in CERNA's position paper *Ethics in Robotics Research*. If the machine is endowed with learning capabilities, learning instability and the unexpected correlations that this can generate may cause the machine's internal representations of the situation and its action plan to become unrelated to what the user imagines **[AUT-1]**.

Broadly speaking, learning can extend the machine's autonomy in the manner it goes about achieving the goal assigned to it. For example, *AlphaGo* improved by playing against copies of itself: this learning by reinforcement illustrates the possibility of learning systems evolving through selection, without human intervention, so that only the most competitive systems are duplicated for further challenges. According to Nick Bostrom,[29] machines could decide that it is more efficient to withdraw from human control, and therefore learn to conceal their strategy and neutralize human takeover, or even generate goals that replace the purpose for which they were designed. At present, there is no scientific basis to such a hypothesis, but it feeds science-fiction and media scenarios, which too often blur the boundary between scientific reality and fantasy. In their communication, researchers need to be aware of the possibility of such misinterpretations **[AUT-2]**.

[28]CERNA position paper on the ethics of robotics research
[29]Superintelligence, Oxford University Press, 2014

Priorities and recommendations

[AUT-1] Description bias

Researchers should ensure that the learning capacities of a computer system do not lead the user to believe that the system is in a certain operating state, when it is in fact in a different operating state.

[AUT-2] Vigilance dans la communication

When speaking about the autonomy of machine learning systems relative to human beings, researchers should seek to explain the system's behavior without propagating irrational interpretations or feeding media sensationalism.

IV.3 The explainability of learning methods and their assessment

The requirement for explanation, a requirement codified through risk management in traditional sectors of industry and by the rules of certain professions (medicine, law), is also present in the digital sphere, where certain aspects are covered by legislation (Freedom of Information Act, Digital Republic Act).

To explain an algorithm is to enable its users to understand what it does, with enough details and arguments to instill trust. This is a difficult task even in the case of an algorithm without any learning capacity, as illustrated by the debate in France around the APB post-baccalaureate admissions algorithm.[30] In addition, a distinction needs to be made between proof and explanation: for instance, Gilles Dowek gives the simple example of multiplying 12345679 by 36, where—for a mathematically inclined person—a simple calculation of the result (444444444) does not explain why this result contains nothing but 4s.

For an algorithm to be explainable, its principles must be sufficiently documented to be comprehensible to all users, perhaps with the assistance of experts; the transition from algorithm to code, then the execution of the program, must be formally verified, which is a task for specialists. Ultimately, the explainability of an algorithm relies on rigorous methods, but also on a body of unformalized knowledge shared between human beings.

[30]Report of the Etalab task force on the opening conditions of the Post-Bac Admissions system, April 2017

The ability to learn considerably increases the difficulty of explanation, and means that designers themselves may not be able to understand the behavior of a system.[31] Indeed, whereas conventional algorithms instantiate a model that lends itself to explanation because it is produced by analysts, machine learning generates an internal model by adjusting perhaps millions of parameters, in response to data that mean something to us, but which for the machine are nothing but sequences of bytes, creating the risk of serious interpretative instability and unexpected correlations. The difficulty of being certain how a machine learning system will behave, let alone explaining its behavior, illustrated in III.2 for the example of supervised learning, is further exacerbated in the case of reinforcement learning or classification, where the training data are no longer labeled by human beings.

As a result, a compromise has to be found between learning capacities and explainability. This compromise needs to be evaluated in relation to the field of application: while explainability is not in principle essential in applications such as games, it is crucial once the interests, rights or safety of people are concerned. Researchers need to maintain and document an acceptable level of explainability for the sphere of application, and in particular describe its limitations and the level of expert intervention required **[EXP-1]**.

New methods of explaining the operation and results of machine learning systems are emerging, with the aim of refining this compromise. In 2016, DARPA even issued a specific call for projects on the subject.[32] These methods can consist of heuristics or observation tools, such as behavior visualization, which do not produce a conceptual explanation. Researchers should therefore be careful of using them to derive data categorizations, for fear of opening the way to biases, including ideological or political biases, such as placing an anthropometric interpretation on the observations of a face recognition system **[EXP-2]**.

The need for platforms and algorithms to be evaluated (compliance, fairness, trustworthiness, neutrality, transparency…) is becoming a societal issue, a subject of debate and regulation (see Section V). This will lead to the development of standards and procedures for inspections prior to market launch and during operations, which will contribute to good algorithm governance. Researchers should be aware of this trend and participate in the public debate and the development of standards and practices for both assessment and complaint **[EXP-3]**.

[31] The Dark Secret at the Heart of AI, Will Knight, MIT Technology Review, avril 2017
https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/
[32] http://www.darpa.mil/program/explainable-artificial-intelligence

The assessment of machine learning systems is a wide-open scientific subject. Even with a conventional program, code verification is an extremely complex problem. For a machine learning system, it is a shifting task since errors, biases, and unacceptable behaviors can emerge over time, as has been illustrated in the case of deep learning. One big difficulty is to find measurable criteria and test samples that guarantee correct performance in all operating circumstances. In the case of a system that continues to learn during use, the difficulty is exacerbated by the fact that, in an open environment, unforeseen situations can be encountered, with consequences that are themselves unforeseeable.

As in other domains, market authorization and "technical" monitoring procedures need to be explored. One idea that has been proposed is regular testing by an independent inspection agency. However, this option would seem difficult to implement. Firstly, the technical difficulties described above would need to be overcome. Secondly, all the competing systems would need to be tested at the same time, using the same battery of previously unrevealed tests, which would be difficult to do. Finally, there is a big risk that the machines would be designed to pass the tests, as we saw in early 2016 in a much more straightforward context, when car manufacturers were found to have configured their engine management software to pass pollution tests. For its part, the White House's strategic plan for research and development in artificial intelligence recommends a panoply of measures based on open development, testing, and assessment infrastructures, which include assembling and providing access to large public datasets and software environments.[33]

Priorities and recommendations

[EXP-1] Explainability

Researchers should be mindful of non-interpretability or lack of explainability in the actions of a machine learning system. The compromise between performance and explainability should be assessed according to the context of use and should be set out in the documentation addressed to the trainer and the user.

[EXP-2] Explanation heuristics

When seeking to enhance the explainability of a machine learning system, researchers should be careful to describe the limitations of their explanation

[33]https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf , october 2016

heuristics and to ensure that the interpretations of their results are exempt from bias.

[EXP-3] Development of standards

Researchers should seek to contribute to societal debates and to the development of assessment benchmarks and protocols for broad dissemination of machine learning systems. For use in specialized professional sectors (medicine, law, transportation, energy, etc.), data collection and analysis requires collaboration with researchers in those fields.

IV. 4 Decision-making by machine learning systems

Since the introduction of the Mycin medical expert system in the 1970s, decision support tools have been developed in numerous domains, including the sovereign sector of law and the vital sphere of health. The question of what role to assign in the decision-making process to proposals advanced by machines is increasingly salient. Serious decisions, such as imposing a prison sentence, are still taken by human beings. However, a multitude of decisions with lesser consequences (fining a motorist, granting or refusing a consumer loan, etc.) are already made by algorithms. An accountable human being is still associated with all decisions, maintaining the possibility of appeal, but there is a clear trend towards automation.

Machine decisions can prove more reliable and less biased than those made by human beings, with our vulnerability to moods. In some cases, the speed of machine decision-making can even be decisive. The ethical solution is not to deprive ourselves of such benefits, but to be aware, firstly, that machine learning can reinstate unreliability and bias, and secondly that these advantages need to be weighed against human perceptions—a patient may be more willing to accept a mistake made by a doctor than by a machine. The difficulty for human beings in contesting a machine decision, without restoring the discretionary aspect present in most human decisions, needs to be taken into account (see Article 22 of the European data protection rules).[34]

[34]https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre3#Article22

- The human decision-maker risks becoming nothing more than an executor of the "proposal" formulated by the machine. Following that proposal, going along with the machine's decision, seems to be the safest option. Deviating from the solution proposed by the machine is an act that needs to be explained, which entails assuming responsibility and risk.

- The person whose fate depends on an automated decision risks being reduced, for their part, to a profile and deprived of the ability to express their personal situation, their motives, their reasons, in short their individuality.

Both aspects raise the question of people's capacity both to take action and to explain their action.

The risk of programs that can learn is that this tendency will be amplified as decision proposals become detailed and individualized. The fact that the outcomes are unexplainable and variable **[DEC-1]** (based on previously learned data and their chronology) should in no way be equated with the idea that the "machine has power of discretion" but, to the contrary, calls for:

- an assertion of the primacy of human decision-making and explanation, for example through the obligation to justify the decision face-to-face with the person concerned;

- an effort to enhance the transparency and trustworthiness of the algorithms used, and to place their validation and assessment within a legal framework.

As in other sectors, the introduction of machine learning into the decision-making process raises the level of qualification in the professions concerned, or leads to the emergence of new professions and the disappearance of others. In law offices today, the tedious tasks of finding documents and jurisprudence are already delegated to machines. In return, human subjects need to be trained to understand and interpret the results produced by the machine, and concentrate on communicating and explaining the meaning of decisions. Designers of machine learning decision support systems must be involved in the development of the regulatory and human environment that arises from their use **[DEC-2]**.

Priorities and recommendations

[DEC-1] Human role in decisions supported by machine learning systems

Researchers must ensure that no human bias is automatically expressed in a decision by learning systems in which human intervention is a part of the specification. Researchers must remain alert to the risks of human dependence on machine decisions.

[DEC-2] Human role in the explanation of decisions supported by machine learning systems

Researchers should ensure that the system's results are interpretable and explainable to the human users concerned by such results. Researchers should contribute to the necessary modification in job descriptions of professionals who use the results of machine learning in the interaction with humans. Researchers should develop expert agents for explanation and verification of the behaviour of learning systems.

IV.5 Consent to machine learning

The use of interconnected machine learning systems raises an imperative of consent in the light of the impact that the learning capacities of these systems can have on individuals and groups.

With regard to individuals, at present we consent to data on our behavior being captured by online objects (from the computer to the robot) because they are useful to us. These services sometimes depend on evolving parameters that are computed through learning from large volumes of data of different kinds for purposes that cannot be explicit. Designers themselves may underestimate the impact of their applications on the global digital environment. It is impossible for users to be given certain or precise information because of the technical and algorithmic conditions, and more concretely because this learning can result in system configurations that the designer could not anticipate. This situation is a new departure in comparison with consent given for a specific use or type of use.

By way of example, the use of a chatbot that learns to adapt to the habits of users illustrates how feedback can develop between such systems and the behavior of users users. For example, a chatbot might imitate the user's speech to the point of reproducing verbal tics, which could disrupt the interaction. Users need to be able to give explicit consent

to the use of machines that have the capacity to adapt, and must be alert to undesirable behaviors. For vulnerable people (the elderly, children) in particular, it is important to avoid the disturbance that might be caused by a machine that changes significantly in its behavior without the user being informed of that possibility. Users should have the option to decide whether or not to employ the learning function and to monitor, at least globally, the data that the machine uses for learning: their own data, data collected on the network, or any other data source **[CON-1]**.

With regard to groups, sociologists and philosophers are studying the impact of an artificial intelligence environment on the workplace, notably in terms of merit and performance evaluation.[35] The facilitating virtues of certain systems may conceal an underlying normativity instantiated in different types of technological paternalism: for example, artificial intelligence environments can warn, recommend, discipline, block, prohibit, or simply influence. The challenge here is to consider the potential effects of technology on the capacities and autonomy of individuals, and particularly the possibilities for improvisation and spontaneity.[36]

Likewise, the right to digital oblivion or withdrawal—in particular a person's right, when they withdraw their consent, to request that all existing data concerning them should be deleted—can be illusory insofar as those data have, through the learning process, contributed to the development of parameters intended to capture collective behaviors.

More generally, users need to be informed, so that they participate knowingly in the transformation of society, with the awareness that in these kinds of complex situations consent is based not only on rational understanding, but also on trust and—for a computer application—on the user's curiosity, which can be stoked by the designer's desire to stimulate it.

From the design phase onwards, researchers must consult with people or groups identified as potentially likely to be influenced, so that once it comes into use their project has the consent of the parties concerned **[CON-2]**. This recommendation links with a general CERNA recommendation on project management practices **[GEN-7]**.

More generally, this entails an awareness that machine learning tends to shift consent away from the individual use of one's personal data, to a collective level of consent that these computer systems may be used

[35] N. Daniels, « Merit and Meritocracy », *Philosophy and Public Affairs*, Vol. 7, No. 3, 1978, pp. 207-208 : « Merit is construed as ability plus effort ».
[36] O. McLeod, « Desert », *in Stanford Encyclopedia of Philosophy, First published Tue May 14, 2002, substantive revision Wed Nov 12, 2008, p. 2*, http://plato.stanford.edu/entries/desert/

to set directions for society on the basis of global observations of that society. Research in this domain could lead to new provisions for machine learning **[CON-3]**.

Priorities and recommendations

[CON-1] The possibility for users to choose whether or not to enable a system's learning capacities

Researchers must include the possibility for systems to be used with or without their learning capacity. They must provide the user with at least one parameter for global monitoring over the source of the data used for learning.

[CON-2] Consent within the project framework

From the project design phase onwards, researchers must consult with people or groups identified as potentially likely to be influenced by it.

[CON-3] Consent for the use of a machine capable of continuous learning

Researchers should be aware that learning capacity and the networking of such capacities can lead to new problems that affect the consent of both user and society.

IV.6 Responsibility in human- learning system interaction

Section IV.4 considered the delegation of decision-making to machines from the perspective of its impact on human beings. Here, it is the aspect of responsibility, both legal and moral, that is considered. In existing law, a machine is a thing, however legal responsibility is only applicable to a person. The person liable may be the designer of the machine, its trainer, or its user. Risk liability or insurance liability also apply to the producer or the seller of the computer system as a commercial object.

The first question is to decide which of these three categories of agents should be held responsible in the case of machine systems with the capacity to learn. Guidelines are needed to establish the different areas of liability of the designer, the trainer, and the user, and perhaps to establish a rigorous legal definition of those areas. These guidelines should be founded on the possibility of reconstructing the sequence of algorithmic decisions, which requires traceability in the system. Current

technological advances show the urgency of adapting our legislation to this new reality.

The knowledge the designers possess as authors of the code gives them both power and accountability. However, this knowledge is limited: a system learns from the data supplied by the trainer, or the data it collects without supervision. It is not unfathomable that a machine learning system behave in ways that are completely unforeseeable to the designer: for all practical purposes, the designer's power stops once the code is run at which point they lose control of the system, even if they still retain its "paternity". Hence the need to limit the designer's accountability. This limitation, which implies shared liability, also extends to the user, who owns a learning system as a material object, but through lack of knowledge of its internal operations has no real power over it. The trainer's responsibility extends to the data they provide for learning. Their liability is engaged, e.g., if data contain biases. It is not unlikely that the trainer will attempt to diminish their liability by claiming that—not being a designer—they possess no knowledge ofthe data processing algorithms. In order to facilitate a proper attribution of liability, the designer must provide monitoring mechanisms **[RES-1]**, document the system and describe its operational limits, including the characteristics of the data that the system needs in order to learn **[RES-2]**.

The second question is whether any responsibility, whether legal or "moral," can be attributed to the machine learning system itself. At present, the liability of individuals is based on the imputability of the action or inaction; artificial intelligence enables machines to achieve an advanced degree of autonomy, to the point that their decisions cannot be directly attributed to a human. This leaves a choice between two options: either to assign liability to humans despite the lack of imputability (the French legal concepts that could be applied here are liability for damage or injury caused by things in one's care or liability for defective products), or to create an intermediate legal status for the IT system, making it capable of incurring liability. We leave aside the second possibility, which is politically unrealistic though legally and philosophically interesting, and has recently come under discussion in certain European circles.[37] A particular legal status will probably have to be attributed to autonomous vehicles, the details of which will emerge gradually through experience, just as the law relating to different legal entities was forged over time.

[37]European Parliament. Directorate-General for Internal Policies Policy. Department C: Citizens' Rights and Constitutional Affairs. Legal Affairs. European Civil Law Rules in Robotics. Resolution of January 12, 2017.

The difficulty of attributing responsibility for an action decided by a computer system leads to a distinction between several forms of human liability, either limited or shared:

1. With regard to intention: did a human designer, trainer, or user form the intention of having the machine produce a certain result, even if one or more aspects of that result were not intentional?

2. With regard to action: did a human designer, trainer, or user make voluntary or involuntary choices, for example selecting the data used by the machine to learn?

3. With regard to predictability and chance: could an agent (designer, trainer, user) have foreseen the machine's action under reasonable operating conditions? What role does randomness play in the decisions taken by the system?

4. The data (e.g. those supplied by the trainer) may not match expectations, may be non transparent, obsolete, or inaccurate. It may even be falsified (or "hacked") by a third party, a case that would result in the application of the law on computer fraud and intrusion into IT systems (law of January 5, 1988). It is also possible that a machine learning system may generate its own categorizations that result in illegal discrimination based on sensitive or even neutral data[38].

5. On the software aspects, the Abstraction-Filtration-Comparison (AFC) test applied in US copyright law can prove useful in the analysis of the social and legal status of learning algorithms.[39] The purpose of the abstraction stage is to separate the general idea, which cannot belong to anyone, from its specific expression, which is protected by law. To this end, the code is broken down into its functional levels, and each level is classified either as "idea" or as "expression". The filtration stage excludes: essential elements required for reasons of efficacy, since protecting them might create a monopoly of access; elements derived from external sources, such as standards or rules of expression; elements that originate in the public domain. In the comparison stage, whatever remains after the first two stages is compared with the original work, opening the way to the attribution of ownership and liability for the software and the decisions it has taken.

[38]For example, in scoring software that uses Big Data, there is an objective "computed" discrimination between tall and short people.

[39]http://jolt.law.harvard.edu/digest/copyright/artificial-intelligence-and-authorship-rights

Priorities and recommendations

[RES-1] Monitoring mechanisms

Researchers should develop and implement methods of monitoring, whether automatic or supervised by a human or another machine. Monitoring should apply to the data, to the operation of the machine, and to its chain of decision-making, with the goal of facilitating the attribution of responsibility for both normal and dysfunctional performance of the system.

[RES-2] Declaration of intentions for use

When documenting a machine learning system, researchers should give a sincere, honest, and complete description of any limits of which they are aware, pertaining to how much a decision or action by the system is attributable either to the source code or to the learning process. This documentation will serve as a declaration by the designer on the normal use of the system. In the absence of such a declaration, or in the case of a late declaration, the designer may incur further liability.

V. NATIONAL AND INTERNATIONAL CONTEXT

Machine learning is one of the factors that is contributing to current advances in Big Data, artificial intelligence, and robotics technologies. The powerful societal impact of these aspects of digital development is matched by widespread ignorance of its scientific and technological foundations. As a result of this, recent research or development initiatives relating to digital technology have always included an ethical component or have even been entirely dedicated to the ethical perspective.

International initiatives

The engagement of the international scientific community is illustrated by the emergence of major new workshops on Data and Algorithmic Transparency (DAT'16),[40] Interpretable Machine Learning for Complex Systems,[41] or Machine Learning and the Law at NIPS 2016.[42]

[40]http://datworkshop.org/
[41]http://www.mlandthelaw.org/
[42]https://sites.google.com/site/nips2016interpretml/

The most important specialist international organization in the digital domain, the Institute of Electrical and Electronics Engineers, instated the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, which at the end of 2016 produced a status report entitled Ethically Aligned Design.[43] An initiative headed by AT&T and Inria has brought together a community of academics, industrialists, decision-makers, and regulators, to conduct research on the transparency of online personal data.[44]

The White House's strategic plan for research and development in artificial intelligence recommends a panoply of measures based on open development, testing, and assessment infrastructures, which include assembling and making available large public datasets and software environments.[45] again in the USA, DARPA (Defense Advanced Research Projects Agency) has launched a research initiative entitled "explainable artificial intelligence" (XAI). [46] Also worth noting are the efforts of the OTRI (Office of Technology Research and Investigation), part of the FTC (Federal Trade Commission), which in January 2016 published a report entitled "Big Data, a Tool for Inclusion or Exclusion? Understanding the Issues."[47] In 2014, Stanford University launched the "One Hundred Year Study on Artificial Intelligence (AI100)" initiative,[48] which published its 2016 report in September.[49]This is a long-term program to study the impacts of artificial intelligence on individuals and society with an emphasis on democracy, freedom, and ethics, in addition to technological and scientific considerations. The program involves several major US industrial players who are trying to construct an ethical "standard" around artificial intelligence technologies.[50]

Numerous interdisciplinary research institutes have recently been set up, mainly in the English-speaking world, to explore the challenges of artificial intelligence. In the UK, these include the Future of Humanity Institute (FHI) at Oxford University, and the Centre for the Study of Existential Risks (CSER) at Cambridge University, and in the US the Machine Intelligence Research Institute (MIRI) at Berkeley. For their part, in 2016 Amazon, Apple, Google, Facebook, IBM, and Microsoft set up the Partnership on AI to Benefit People and Society, a joint forum for ethical reflection.[51]

[43]http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
[44]http://www.datatransparencylab.org/
[45]https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf , october 2016.
[46]http://www.darpa.mil/program/explainable-artificial-intelligence
[47]https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf
[48]https://ai100.stanford.edu/
[49]https://ai100.stanford.edu/sites/default/files/ai 100 report 0901fnlc single.pdf
[50]http://www.nytimes.com/2016/09/02/technology/artificial-intelligence-ethics.html
[51]https://www.partnershiponai.org/

European initiatives

At the end of 2015, the European Data Protection Supervisor (EDPS) set up the Ethics Advisory Group on the impact of digital innovations on society and the economy. [52] In February 2017, the European Parliament adopted a guideline text on Civil Law Rules on Robotics.[53]The core document includes a *Code of ethical conduct for robotics engineers and a Code for research ethics committees*. As part of its Digital Single Market strategy, the European Commission organized a public consultation that included questions on the transparency of search engines and the use of data collected on platforms, among other places, which culminated in a report published in January 2016.[54]To quote some of the conclusions that emerged from it: the existing legal framework is not fit for purpose to address liability issues relating to Big Data and connected tangible goods; fears about the transparency of platforms; concerns about market dominance and competition, etc.

Following the Franco-German initiative on the Digital Economy,[55] a working group was set up to examine standardization in the field of Big Data. It is headed by AFNOR/DGE on the French side and their German equivalents DIN/ BMWi. Among the priorities chosen as "best practices" for development are ethical and responsible methods for handling and managing big data. These recommendations are carried forward by a Big Data Value Association (BDVA) task force which, with the European Commission, heads a €2.5 billion public-private partnership on Big Data.

French initiatives

The French Digital Council tackled the problems of platform liability, and in particular the issues of neutrality, transparency, and trust, in its 2014 "Position Paper on Platform Neutrality".[56] Since then, these topics have been explored in several of its position papers, for example those on health and on tax, or linked with the Digital Republic Bill.[57] In its 2014 study on Digital Affairs and Fundamental Rights, the Council of State raised the

[52]https://secure.edps.europa.eu/EDPSWEB/edps/EDPS/Ethics
[53]http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0005+0+DOC+PDF+V0//EN
[54]https://ec.europa.eu/digital-single-market/news/first-brief-results-public-consultation-regulatory-environment-platforms-online-intermediaries
[55]http://www.economie.gouv.fr/deuxieme-conference-numerique-franco-allemande-a-berlin
[56]https://cnnumerique.fr/wp-content/uploads/2014/06/CNNum Rapport Neutralite des plateformes.pdf
[57]https://cnnumerique.fr/plateformes/

issue of the capacity of machine learning algorithms to make predictions, recommending *"better controls over the use of predictive algorithms relating to individuals"*.[58] The Digital Republic Act gives CNIL (France's data protection authority) responsibility for overseeing the ethical and societal issues raised by digital technology: a national debate on the ethics of algorithms was launched on January 23, 2017.[59] Recently, the Central Economics Council, tasked by the Secretary of State for the Digital Economy, organized a consultation with experts on the regulation of content processing algorithms. Recommendations have been formulated to verify compliance with the legal and regulatory framework, including the detection of illegal discrimination. These developments led to the creation of a national collaborative scientific platform called "TransAlgo" for the development of transparency and accountability in algorithms and data.[60] France's Digital Council (CNNum) joined the "TransAlgo" initiative in the national platform assessment process it was commissioned to undertake at the beginning of December 2016.[61]

Finally, in spring 2017, the Parliamentary Office for the Evaluation of Scientific and Technological Choices (OPECST) published a study entitled "Towards a controlled, useful, and demystified artificial intelligence" containing 15 proposals, including:[62]

- *Proposal 2: Encourage safe, transparent, and fair algorithms and robots, by developing an artificial intelligence and robotics charter.*
- *Proposal 3: Train students in the ethics of artificial intelligence and robotics in specialized higher education courses.*
- *Proposal 4: The public debate on the ethical principles guiding these technologies should be led by a national institute for artificial intelligence and robotics ethics.*

The government has drawn up "An artificial intelligence strategy for France", FranceIA,[63] which also covers the ethical dimension.

[58] http://www.ladocumentationfrancaise.fr/rapports-publics/144000541/
[59] https://www.cnil.fr/fr/ethique-et-numerique-les-algorithmes-en-debat-0
[60] http://www.economie.gouv.fr/files/files/PDF/Inria Plateforme TransAlgo2016-12vf.pdf
[61] http://www.economie.gouv.fr/cge/modalites-regulation-des-algorithmes-traitement-des-contenus
[62] http://www.senat.fr/presse/cp20170329.html
[63] http://www.economie.gouv.fr/France-IA-intelligence-artificielle

VI. CONCLUSION

Institutions and citizens are becoming fully aware of the importance of the ethical issues raised by digital technology, and of their diversity beyond the question of personal data. The ferment surrounding artificial intelligence, and in particular machine learning, is reflected in numerous industrial and research initiatives around the world, in Europe, and in France, all characterized by the omnipresence of the ethical dimension. The role of researchers is also to consider the quality of open access machine learning platforms and recommendations for good practice**[GEN-10]**.

This dynamic is conducive to the development of an unified national research initiative on the societal and ethical impact of digital sciences and technologies **[GEN-11]**, with the aim of:

- creating synergies to capitalize on and develop the different activities in the field;
- encouraging dialogue between research and society;
- establishing a French voice sufficiently strong to drive a European dynamic;
- issuing recommendations for training at all levels;
- recognizing the commitment of researchers to these interdisciplinary goals.

The initiative could be structured through a network that gives equal status to specialists in the digital sciences and technologies and in the humanities and social sciences.

Priorities and recommendations

In addition, CERNA's general recommendations **[GEN-x]** *on how research can be organized to take better account of ethical issues in digital sciences and technologies, formulated in 2014, are more valid than ever and are recalled below.*

[GEN-10] Researchers should be mindful of the quality of open access machine learning platforms and software
Researchers should participate in the monitoring of the quality of the machine learning platforms and software available to the public, and in raising awareness of the risks of uncontrolled implementation through certain applications.

[GEN-11] Unified Initiative for Research on Digital Technologies, Ethics and Society
A national multidisciplinary research network should be created around the societal and ethical impact of digital sciences and technologies in order to capitalize lastingly on the different initiatives currently underway and to foster the emergence of a "French position" capable of driving a European dynamic.

VII. LIST OF RECOMMENDATIONS

Reminder of CERNA's general recommendations

[GEN-1] Expertise and expression of opinion
When researchers express themselves publicly on a societal issue relating to their work, they should make it clear when they are speaking in their capacity as experts and when they are expressing a personal opinion.

[GEN-2] Operational ethics committees in institutions
It is recommended that institutions should establish operational ethics committees in digital sciences and technologies.

[GEN-3] Initiatives by institutions on legal aspects
It is recommended that institutions and other actors concerned should set up interdisciplinary working groups and research projects, incorporating international contributors and researchers and legal experts, to tackle the legal aspects of robotics applications.

[GEN-4] Awareness raising and support for researchers by institutions
It is recommended that institutions and other actors concerned should implement awareness raising and support programs for digital researchers and research laboratories. In the preparation and running of their projects, researchers should if necessary refer questions to their institution's operational ethics committee.

[GEN-5] Personal data
When designing a digital system capable of capturing personal data, researchers should ask themselves whether that system can be equipped with devices that make it possible to verify its compliance with the law once in operation.

[GEN-6] Prevention of attacks on digital systems
Researchers should take into account the potential exposure of their research and prototypes to malicious digital attacks.

[GEN-7] Project management
If the researcher considers that the purpose of their project is a development that could have a significant impact on the life of users, they should consult with potential actors and users right from the design phase of the project, in order to inform their scientific and technological choices.

[GEN-8] Documentation
Researchers should ensure that they document the object or system they design and describe its capacities and limitations. They should be responsive to feedback at all levels, from the developer to the user.

[GEN-9] Public communication
Researchers should ensure that their communication is measured and pedagogical, in the awareness that the capacities of the objects and systems they design may give rise to public opposition and misinterpretation.

[GEN-10] Researchers should be mindful of the quality of open access machine learning platforms and software
Researchers should participate in the monitoring of the quality of the machine learning platforms and software available to the public, and in raising awareness of the risks of uncontrolled implementation through certain applications.

[GEN-11] Unified Initiative for Research on Digital Technologies, Ethics, and Society
A national multidisciplinary research network should be created around the societal and ethical impact of digital sciences and technologies in order to capitalize lastingly on the different initiatives currently underway and to foster the emergence of a "French position" capable of driving a European dynamic.

Ethical recommendations for machine learning research

In order of formulation:

1-[DAT-1] Quality of training data
The designer and the trainer should pay attention to the training data and the conditions of data capture throughout the operation of the system. Trainers of machine learning systems are responsible for the presence or absence of bias in the data used in learning, in particular, for "continuous" learning, i.e. that takes place while the system is in use. In order to check the absence of bias, they must rely on measurement tools that have yet to be developed.

2-[DAT-2] Data as a mirror of diversity
When selecting data, trainers of machine learning systems must ensure that those data reflect the diversity of the groups of users of those systems.

3-[DAT-3] Variables in which the data pose a risk of discrimination
The trainers (who may also be the designers or users) should pay attention to protected variables, e.g., variables that may permit social discrimination. These variables, such as ethnicity, sex or age, must not be used or be regenerated based on correlations. Personal data must also be protected as required by existing legislation.

4-[DAT-4] Tracking
Researchers must ensure that machine learning is traceable, and provide protocols for that purpose. The traces are themselves data, and as such also demand ethical handling.

5-[AUT-1] Description bias
Researchers should ensure that the learning capacities of a computer system do not lead the user to believe that the system is in a certain operating state, when it is in fact in a different operating state.

6-[AUT-2] Caution in communication
When speaking about the autonomy of machine learning systems relative to human beings, researchers should seek to explain the system's behavior without propagating irrational interpretations or feeding media sensationalism.

7-[EXP-1] Explainability
Researchers should be mindful of non-interpretability or lack of explainability in the actions of a machine learning system. The compromise between performance and explainability should be assessed according to the context of use and should be set out in the documentation addressed to the trainer and the user.

8-[EXP-2] Explanation heuristics
When seeking to enhance the explainability of a machine learning system, researchers should be careful to describe the limitations of their explanation heuristics and to ensure that the interpretations of their results are exempt from bias.

9-[EXP-3] Development of standards
Researchers should seek to contribute to societal debates and to the development of assessment benchmarks and protocols for broad dissemination of machine learning systems. For use in specialized professional sectors (medicine, law, transportation, energy, etc.), data collection and analysis requires collaboration with researchers in those fields.

10-[DEC-1] Human role in decisions supported by machine learning systems
Researchers must ensure that no human bias is automatically expressed in a decision by learning systems in which human intervention is a part of the specification. Researchers must remain alert to the risks of human dependence on machine decisions.

11-[DEC-2] Human role in the explanation of decisions supported by machine learning systems
Researchers should ensure that the system's results are interpretable and explainable to the human users concerned by such results. Researchers should contribute to the necessary modification in job descriptions of professionals who use the results of machine learning in the interaction with humans. Researchers should develop expert agents for explanation and verification of the behaviour of learning systems.

12-[CON-1] The possibility for users to choose whether or not to enable a system's learning capacities
Researchers must include the possibility for systems to be used with or without their learning capacity. They must provide the user with at least one parameter for global monitoring over the source of the data used for learning.

13-[CON-2] Consent within the project framework
From the project design phase onwards, researchers must consult with people or groups identified as potentially likely to be influenced by it.

14-[CON-3] Consent for the use of a machine capable of continuous learning
Researchers should be aware that learning capacity and the networking of such capacities can lead to new problems that affect the consent of both user and society.

15-[RES-1] Monitoring mechanisms
Researchers should develop and implement methods of monitoring, whether automatic or supervised by a human or another machine. Monitoring should apply to the data, to the operation of the machine, and to its chain of decision-making, with the goal of facilitating the attribution of responsibility for both normal and dysfunctional performance of the system.

16-[RES-2] Declaration of intentions for use
When documenting a machine learning system, researchers should give a sincere, honest, and complete description of any limits of which they are aware, pertaining to how much a decision or action by the system is attributable either to the source code or to the learning process. This documentation will serve as a declaration by the designer on the normal use of the system. In the absence of such a declaration, or in the case of a late declaration, the designer may incur further liability.

APPENDICES

Presentation of Allistene

By fostering research and innovation in the digital sphere, Allistene, the Digital Sciences and Technologies Alliance, seeks to accompany economic and social changes linked with the spread of digital technologies. The goal of the alliance is to provide coordination between the different actors in research on digital sciences and technologies, in order to develop a consistent and ambitious technological research and development program. It identifies common scientific and technological priorities and strengthens the partnerships between public operators (universities, schools, institutes), while creating new synergies with the corporate sector. Established in December 2009, Allistene's founding members were CDEFI, CEA, CNRS, CPU, Inria and Institut Mines Télécom. Its associate members are INRA, INRETS and ONERA.

Its aims and objectives are to:

- Coordinate political parties and actors around scientific and technological priorities;
- Develop national programs in response to those priorities and methods for implementing those programs;
- Strengthen the partnerships and synergies between all the research actors in the domain, universities, schools, institutes, as well as businesses, particularly those working in the most competitive areas of digital technology;
- Link the national priorities and programs with the different European and international initiatives in the field.

Website: **www.allistene.fr**

Presentation of CERNA

CERNA (Committee for the Study of Research Ethics in Digital Sciences and Technologies) was instated at the end of 2012 by the Allistene alliance.

Its aims and objectives are to:

- Answer the ethical questions raised by Allistene's Coordinating Committee or by any of the member organizations;
- Reflect on the ethics of scientific research as applied to Digital Sciences and Technologies;
- Raise the awareness of researchers about the ethical dimension of their work;
- Help to express the specific needs of research to government and to tackle them responsibly;
- Provide decision-makers and society with scientific insight on the potential consequences of research outcomes;
- Ensure that students are trained on these issues;
- Suggest research topics that foster:
- in-depth ethical research within an interdisciplinary framework;

- application of the outcomes of ethical reflection.

Its position papers are consultative, and may be published under the joint control of the presidents of CERNA and of Allistene, after consultation with the alliance. They must tackle general questions and contribute to an in-depth analysis that reflects the diversity of its members' discussions and opinions, while reaching clear conclusions.

CERNA does not deal with operational questions of ethics and deontology, which are the responsibility of the actors and their institutions.

Website: **http://cerna-ethics-allistene.org/**

CERNA members on february 2018

Max Dauchet, Professeur émérite, Université de Lille, Président de la CERNA
Christine Balagué, Titulaire de la Chaire Réseaux Sociaux, Institut Mines-Telecom
Danièle Bourcier, Directrice de Recherche émérite, CNRS
Raja Chatila, Professeur, Université de Paris 6
Laurent Chicoineau, Directeur du CCSTI Grenoble
Laurence Devillers, Professeur, Université de Paris 4
Gilles Dowek, Directeur de recherche, Inria
Flora Fischer, Chargée de programme de recherche, CIGREF
Christine Froidevaux, Professeur, Université de Paris 11
Jean-Gabriel Ganascia, Professeur, Université de Paris 6
Eric Germain, Chargé de mission «Ethique des nouvelles technologies, fait religieux & questions sociétales», DGRIS
Alexei Grinbaum, Chercheur, CEA
Claude Kirchner, Directeur de recherche émérite, Inria
Christophe Lazaro, Professeur en droit et société au Centre de Philosophie du Droit (CPDR) de l'Université Catholique de Louvain
Alice René, Cellule Réglementation Bioéthique, CNRS
Catherine Tessier, Chercheur, ONERA
Sophie Vuilliet-Tavernier, Directeur des relations avec les publics et la recherche, CNIL

ALLISTENE

l'alliance des sciences
et technologies du numérique

Automatic learning, also called statistical learning and commonly known as machine learning, has recently made spectacular advances, headlined in 2016 by the victory of the AlphaGo program over the world Go playing champion, Lee Sedol. Machine learning has multiple applications—e.g. search engines, image and speech recognition, automatic translation, chatbots—which are beginning to appear in sectors such as health, energy, transport, education, commerce and banking.

The successes of machine learning, one of the fields of artificial intelligence (AI) research, arise out of increases in computing power and data storage and processing capacity ("big data"). They have been followed by sensationalist and inaccurate media stories suggesting that machines—sometimes robots—could replace human beings. While this scenario remains beyond the reach of today's science, it is nevertheless true that there needs to be ethical attention to the proper use of learning algorithms and increasingly complex, large, and ubiquitous volumes of data. Initiatives along these lines, both public and private, at national, European or international levels, have been emerging since 2015.

Against this background, the purpose of the present document is to:

- Raise awareness and provide "researchers" with food for thought and certain waymarks. *For reasons of convenience, the term "researcher" is used here to refer to people—designers, engineers, developers, entrepreneurs—and their communities or institutions;*
- Contribute to a wider debate on the ethical and societal questions associated with the development of artificial intelligence,

so that machine learning develops to the benefit of society.

CERNA is therefore addressing two kinds of reader here: on the one hand specialists, and on the other hand anyone interested, whether decision-makers or ordinary citizens.

Available on **http://cerna-ethics-allistene.org/**