numeum

BEHIND THE CODES AND THE DATA

A PRACTICAL GUIDE TO ETHICAL AI

SOMMAIRE

THE INITATIVE

- The ambition of Numeum and its partners has been to translate general **ethical principles** based on existing work, into **practical methods.**
- **7 collective intelligence workshops** involving over **350 participants** were conducted between 10 November and 22 December 2020. They allowed us to gather the contributors' **proposals** and **recommendations** on each topic identified.
- This work was consolidated and formalised over the period from January to April 2021, under the supervision of a **review committee** made up of **academic partners** and different **professionals.**
- **Important note:** the reflection deliberately focused on simple AI systems, making it easier to project onto concrete application cases. Complex AI systems involving fully automated decision-making processes such as autonomous vehicles and robots operating without human intervention were excluded from the work.
- This **Practical guide** has been officially presented to the public authorities. It also forms the basis for the Manifesto for Ethical AI You can find all of this information on the website dedicated to the Ethical AI initiative: ai-ethical.com.
- This document will be completed and **regularly updated** to keep up with **technological developments** or **new requirements** (regulatory, societal, economic, etc.).

Renaud VEDEL, *Coordinator of France's National AI Strategy (CSN-IA)*

Katya LAINÉ, *Director and Chair of Numeum's Artificial Intelligence Committee*

NO RESPONSIBLE DIGITAL TECHNOLOGY WITHOUT ETHICAL AI

Bringing unprecedented progress in all kinds of areas, «Articifical Intelligence (AI) is spreading through our daily lives. But this phenomenal expansion raises the **legitimate question of the trust** we humans can put in these systems.
To answer this question, the solutions developed must respect the fundamental rights defended, in particular, by France and the European Union. To fulfil all its promise, AI therefore has to be ethical: this is one of the essential conditions if it is to bring all the **benefits it promises to as many people as possible.**
But when it comes to putting the theory into practice, the task is not so easy.
So it is with the same determination as they put into moving technology forward that AI specialists have mobilised to devise an operational framework for the creation and dissemination of ethical AI systems.

Numeum has formed a network of partners from different AI worlds – academia, public authorities, companies, the voluntary sector and civil society. The numerous exchanges between them and the sharing of their experience have culminated in the creation of a practical guide that proposes **a useful method for implementing the main ethical principles** when designing, developing and deploying AI solutions.
The guide therefore constitutes a **voluntary code of conduct** for developing the trustworthy AI systems encouraged by the European Commission[1]. Thanks to the credibility of the partners backing the initiative, we believe it has the potential to become the accepted baseline in France and beyond.

This initiative to give AI tangible ethical credentials enhances our approach to developing a more accountable digital world. It constitutes a key step in the anticipation and preparation of compliance with the future general regulation on AI recently proposed by the European Commission.

1. See https://ec.europa.eu/commission/presscorner/detail/fr/QANDA_21_1683

PARTNERS

Clic on a picture to launch the associated video.

Laurence DEVILLERS, *Professor of AI at the Institut DATAIA*

Charles BOUVEYRON, *Director of the Institut 3IA Côte d'Azur*

Roxana RUGINA, *General Secretary of Impact AI*

Sophie VIGER, *General Manager of École 42*

Tawhid CHTIOUI, *Founding President & Dean of Aivancity*

Magali BARNOIN, *Digital, Data & AI facilitator at Telecom Valley*

Antoine TROTET, *Head of the Digital Revolution department at the Grand Est regional authority*

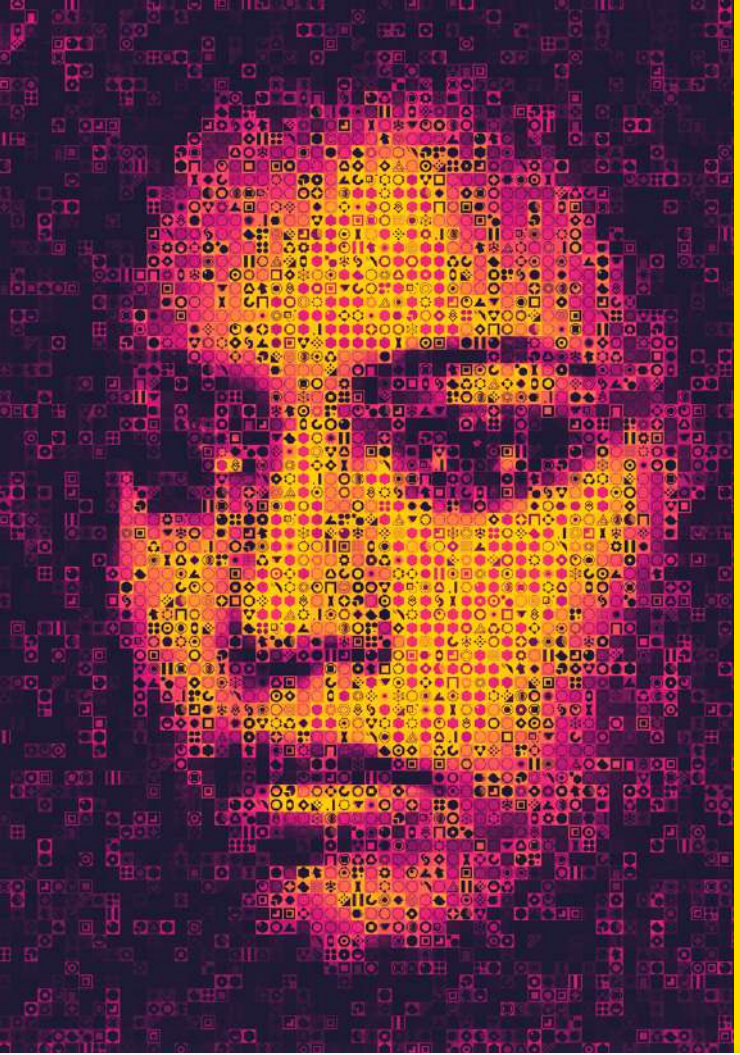
Gaëlle PINSON, *General Manager of Hub France IA*

Nicolas VIALLET, *Chief Operating Officer ANITI, University of Toulouse*

Alexis STEINER, *AI and Digital Project Manager at Grand E-Nov+*

HOW TO USE THE GUIDE

- **Part 1:** this part contains the definitions and the background to the initiative.
- **Part 2:** this part describes the methodology itself in the form of practical recommendations to be put into practice in three steps:
  1. setting up of an AI governance system within the company,
  2. assessment of the ethical risks involved in the project, identification of the areas where efforts should be concentrated using an ethical sensitivity matrix to measure the project's sensitivity to ethical issues,
  3. application of the recommendations based on the areas chosen to focus on, at each stage in the life cycle.
- **Partie 3:** this part contains some examples of projects and use cases to illustrate the implementation of the methodology.



PART 1

FRAMEWORK AND POINTERS

YOUR NOTES:

WHAT'S AT STAKE?

AI promises so much: to help us meet the great societal and environmental challenges ahead, to optimise the way companies operate, to take the strain and discomfort out of certain activities, to do more for our customers, to make diagnoses faster and better, to make our daily lives easier, etc.

And yet these technologies, especially the more recent ones known as deep learning, are raising **questions among the public.**
The aura of mystery that shrouds them probably has something to do with it. But the excesses of some AI systems[1] and the physical injuries caused by others[2] have exacerbated mistrust among users who are now asking themselves whether the designers of AI systems are able to control their creations.

2 - Amazon's AI recruiting tool was found to discriminate against women and Microsoft's Tay chatbot started to spout insults

2 - Uber's self-driving car was responsible for a fatal accident

WHAT'S AT STAKE?

To avoid this wariness turning into a systematic and massive rejection which would obliterate all the potential benefits of these technologies at a stroke, we urgently need to rethink the way we design AI solutions.

From now on we have to want systems that are not only effective, but also ethical, that is to say designed and operated in such a way as to ensure that their use and effects do no harm to either the dignity or the integrity of human beings. They must also respect fundamental ethical values, such as the right to privacy, fair treatment and the freedom to act and make decisions.

Digital professionals whose jobs consist of **designing, developing, operating and decommissioning AI solutions** are the first to be concerned by these issues.

That's why Numeum and its partners are proposing this **practical guide** detailing a methodology for creating AI systems compliant with ethical values and able to meet society's expectations. For this is the price at which we can **build - or rebuild - trust in these technologies**, which is the one condition indispensable to their widespread use.

ETHICAL AI: WHAT EXACTLY ARE WE TALKING ABOUT?

«Legislation does not always keep up with the pace of technological change, sometimes does not correspond to ethical standards or can simply prove to be inappropriate faced with certain issues.
To be trustworthy, AI systems should also be ethical, with care being taken to align them with ethical standards.»
Guidelines on ethics for trustworthy AI of GEHN IA.

An AI system can be qualified as ethical if it is able to preserve the fundamental human rights throughout its life cycle, right to dignity, mental and physical integrity, freedom, autonomy, equal treatment, intimacy and privacy.

These rights are, in theory, protected by current French and European laws and regulations, such as the GDPR, which governs the protection of personal data. An AI system, whether or not it claims to be ethical, of course has to abide by the law or it will be illegal and unable to access the market. But we must go further. Technology is progressing faster than the law. Now, the nature of the AI systems now emerging, especially those using deep learning technologies, and the considerable impact they have on our lives are producing a high concentration of risks from the ethical point of view: discrimination, dehumanisation of social relations, opaque deduction processes, abusive use of personal data, errors caused by cyber attacks, etc.

These risks have to be taken into consideration all the more seriously as trade-offs will inevitably occur to resolve the contradictions between performance objectives and ethical principles.
The situation therefore calls for specific measures and precautions to avoid deliberate or unintended abuses and excesses.

Because, at the end of the day, Humans alone are accountable for the systems they create.

ETHICAL AI: WHAT EXACTLY ARE WE TALKING ABOUT?

«Every person involved in the creation of AI at any step is accountable for considering the system's impact in the world.»
Everyday Ethics for Artificial Intelligence, IBM.

«Reinforcement learning measures should be built not just based on what AI or robots do to achieve an outcome, but also on how AI and robots align with human values to accomplish that particular result»
The Ethics of Code, Sage.

By analysing a number of different texts *[see Bibliography p.13]*, Numeum has endeavoured to characterise ethical AI through a number of principles or qualities that it must possess: An AI system can be qualified as ethical if it is:

1. **Respectful** of personal data.
2. **Unbiased:** it endeavours not to create or reproduce discrimination and favours inclusiveness.
3. **Transparent** in its operation: its conclusions can be explained and it is auditable.
4. **Fair** in its relations with human beings: it only does what is expected of it and it discloses itself to the user.
5. **Controlled:** it remains under the control of humans.
6. **Reliable:** it is secure and robust in the face of cyber attacks.

Each of these qualities reflects a certain number of requirements.
This methodological guide, which is meant to be put to operational use, describes those requirements and proposes state-of-the-art good practices to meet them.
It will then be up to each company to assess the risks inherent in its AI systems from the ethical point of view and to decide the degree of application of the measures recommended based on those risks and the purpose of the system.

Nota bene: *AI can, furthermore, make a positive contribution to society and people's well-being through its purpose (AI for good). This subject is not addressed in this document, which aims to list the practical measures to be implemented to create and deploy ethical AI by design, irrespective of what the system has been designed to do.*

BIBLIOGRAPHY

To provide a framework for the initiative and decide the areas it would work on, Numeum started by consulting a corpus of existing ethics charters and reference publications:

- Algorithmes, contrôle des biais SVP,, a white paper published by the Institut Montaigne (March 2020)
- Montréal Declaration for a Responsible Development of Artificial Intelligence, written by a multidisciplinary, inter-university scientific team from Montréal after a process ofcitizen consultation and co-construction (2018)
- Ethically aligned design (2017) et Establishing standards for ethical technology P70xx (2018, two papers published by the IEEE
- Everyday Ethics for Artificial Intelligence, guide published IBM (2019)
- Ethics guidelines for Trustworthy AI (April 2019) and Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (July 2020), two texts issued by the Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission
- IA responsable : principes, approches et mise en action, a guide issued by Microsoft (2018)
- Building Trust in Human-Centric Artificial Intelligence, communication from the European Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions (COM(2019) 168, April 2019), which contains the guidelines mentioned above
- Responsible AI: A Global Policy Framework, proposed by ITechLaw (2017)
- Rome Call for AI Ethics, a pledge document from the Vatican (February 2020)

- The Ethics of code. Developing AI for business with five core principles, a charter drawn up by publisher Sage (2017)
- Un engagement collectif pour un usage responsable de l'IA, a charter drawn up by Impact AI
- An Artificial Intelligence - A European approach to excellence and trust, a white paper issued by the European Commission (COM(2020) 65, 65) February 2020)
- Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), the first legal framework on AI proposed by the European Commission (April 2021)

PART 2

PRACTICAL RECOMMENDATIONS

YOUR NOTES:

AI GOVERNANCE AT THE HIGHEST LEVEL OF THE COMPANY

The stakes involved in the ethical risks of an AI system are too high to be left to the company's data scientists and other technical specialists. The subject is often highly complex. It requires perspective, especially once you start getting into impact assessments. It may require decisions to be made at the highest level which need to draw on other skills than those of the project team. What's more, it may involve a reputational or legal risk for the company.

That is why the very first recommendation in this guide is to set up a specific governance system at the highest level, which will be responsible for defining and applying the company's policy on AI. Reflecting the company's ethical values and complying with the regulations, this policy will lay down the principles on which all the rest will be based. It will set the framework within which AI projects will be developed, i.e. a set of standardised project management processes and methods intended to secure the treatment of the ethical issues.

This governance system will have the task of defining a method of assessing the ethical risks and ensuring the necessary impact assessments are conducted. Depending on the projects, its role will also consist of arbitrating between any tensions (between performance and ethical and security issues), establishing thresholds and other limits, specifying the criteria of fairness and explainability which will be referred to and examining critical cases.

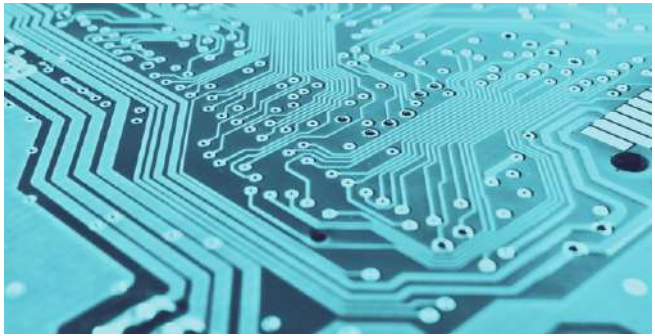
The organisation and scope of this governance will vary from one company to another. In certain cases, the matters of GDPR and NIS (Network and Information System Security) compliance could also be part of its remit.

→ You can refer to the Guide IA digne de confiance (Guide to trustworthy AI) published by Impact AI to set up your AI governance system in your company (Chapter 3). Also see Chapter 4 on self-assessment.
→ Also to assess your AI governance, see: Questionnaire sur la gouvernance des algorithmes d'IA dans le secteur financier (questionnaire on AI governance in the financial sector) used by the ACPR (the French prudential supervision and resolution authority).

AI GOVERNANCE AT THE HIGHEST LEVEL OF THE COMPANY

A few other things AI governance will need to do:

- raise awareness among project teams – data specialists (data scientists, data analysts, etc.) and business specialists – of the ethical challenges of AI as well as the regulatory and security-related issues,
- establish the principle of having a data controller for each project, who will be the company's single point of contact on these issues, both internally and externally,
- set up a process of escalation if an ethical risk arises; this measure requires the creation of a climate of confidence within the company and in its relations with its partners to that anyone in the chain feels able to raise an alert.



→ **An example:** at Microsoft, a special communication channel for alerting the local ethics committee allows anyone who spots an ethics risk in a AI system that is in development to make themselves heard (for example to report that the conclusions of the AI system in question could lead to a substantial service being refused, or cause harm or violate person's rights).

ASSESSING THE ETHICAL RISKS OF A PROJECT

Creating ethical AI by design first and foremost means designing an AI system whose risk of dangerous or prejudicial consequences on an ethical level is reduced to the minimum or eliminated altogether. For each AI project, an assessment of the ethical risk is essential. The analysis will cover the 6 qualities described on page 12 and their associated requirements. It will concern both the risks induced by the use case and those inherent in the technology and those linked to the context of the project.
It will assess the risk in terms of its likelihood and its severity and will compare these with the purpose of the use case.
The study must also take account of the more general risks to the company, its business and its market.

The main types of risks to be assessed:

- Concerning **the use case**
  The major risk is the risk that the result of an AI system could cause to individuals, society, the environment:
  - risk of a result leading to the refusal or limiting of a person's access to a fundamental right (e.g. refusal of access to an essential service);
  - risk of creating an addiction or locking a person into a behaviour;
  - risk of occurrence of unforeseen and potentially uncontrolled cases which could be detrimental to human integrity and/or dignity or impact the environment (this risk can appear in systems with a certain degree of autonomy);
  - risk of discrimination, etc.

- Concerning **the system**
  The main risks are:
  - risk of non-compliance with current data protection regulations (the GDPR in Europe - certain data can turn out be personal data in the eyes of the CNIL, even though they do not seem to directly identify a person, e.g. an IP address),
  - risk of cyber attacks leading to the disclosure of personal information and/or a purpose diversion,
  - risk of opacity of the decision path in the AI system,
  - risk linked to an absence of traceability that prevents the correction of the system and/or the identification of responsibilities, etc.

- Concerning **context of the project**
  The main risk is the risk linked to a failure of AI governance in the company:
  - risk of the teams and/or the company misconstruing what is actually at stake, risk of the lack of a correction and repair process,
  - risk of the absence leads or resource persons, etc.

ASSESSING THE ETHICAL RISKS OF A PROJECT

The results of the assessment will serve to orient the reflections and decide what measures can be taken to reduce the risk. They will allow you to define, among other things, the threshold at which human intervention becomes necessary.

→ A very comprehensive risk assessment tool: Ethics & Algorithms Toolkit. Produced by an American team, it targets the use of AI in the public sector.

The **ethical sensitivity matrix** presented in this guide provides an interpretative framework for the recommendations listed in the pages that follow.
It is intended to help developers to orient their efforts based on the sensitivity of their AI systems to ethical issues.

→ A few extra toolkits and questionnaires that can be used to tangibly measure the ethical impact of your AI system:

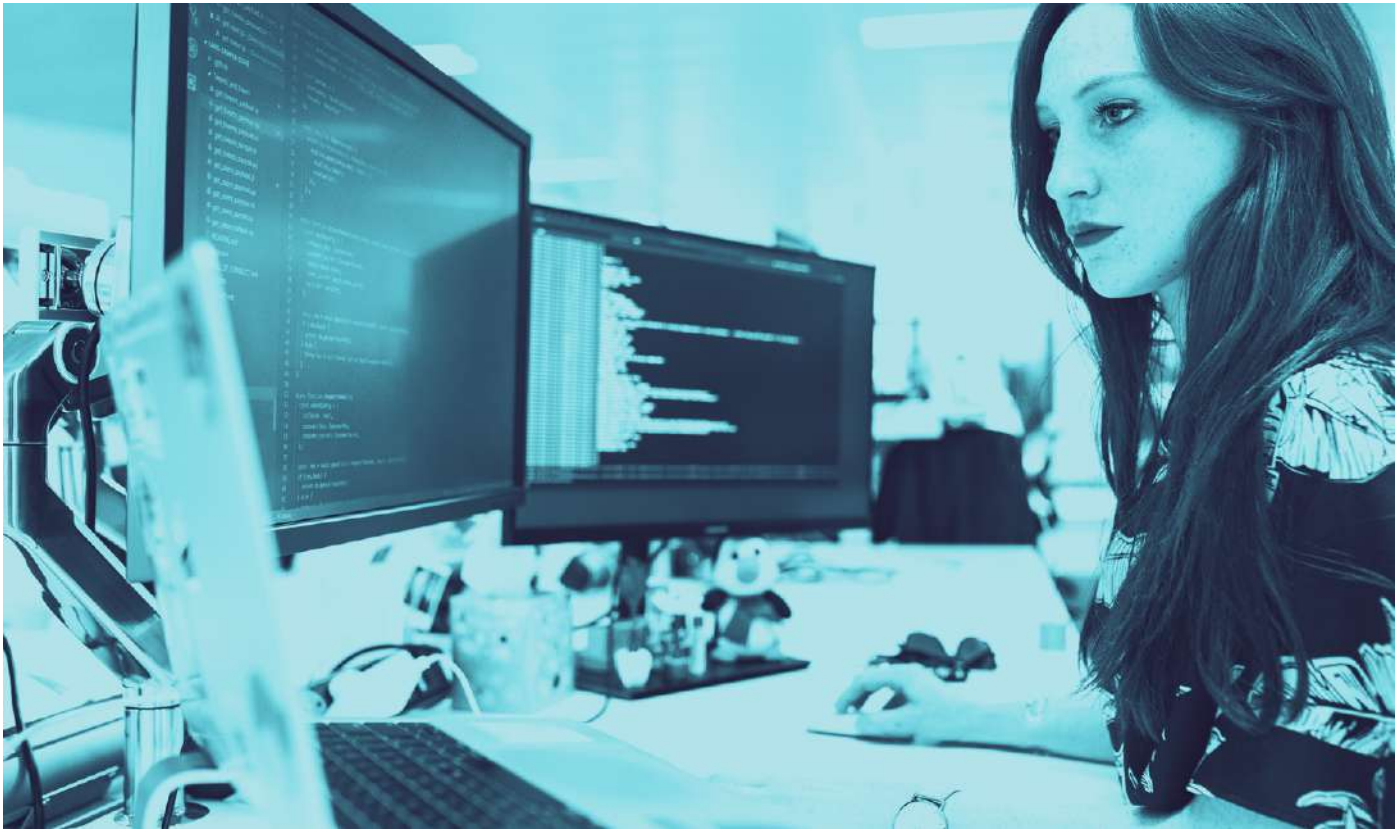
- The Ethics Guidelines of the MAIF, which are included in the Melusine automatic email classification project.
- Référentiel d'évaluation de la maturité d'une organisation, a framework for assessing an organisation's maturity from the Substra Foundation.
- Assessment List for Trustworthy Artificial Intelligence (Altai), the questionnaire of the Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission.
- The Box by AI Ethics Lab, which can be used to gauge, based on different criteria, the strengths and weaknesses of your system from the ethical point of view (read in conjunction with the article *Operationalizing AI ethics principles*).
- Responsible AI Toolkit, an evaluation toolkit offered by PWC.
- Artificial intelligence impact assessment, a very comprehensive guide from independent Dutch platform ECP.

ETHICAL SENSITIVITY MATRIX

Ethical subjects to consider specifically →			Data protection		No bias			Transparency		Fairness		Control	Reliability
↓ System purpose and implementation framework		Applies to the project	Confidentiality of personal data	Controlled and measured use of personal data	Prevention of risks of discrimination	Diversity of the project team	Accessibility of the solution	Explainability of the model and the results	Traceability of data and processes	Reliability of results	Disclosure of the AI	Operation under human control	Robustness and resilience of the solution
The business need	The system automates a decision, or helps to make a decision concerning physical persons	Yes/No			⊗			⊗	⊗	⊗		⊗	⊗
	The system automates the performance of tasks for the user	Yes/No						⊗	⊗	⊗	⊗	⊗	⊗
	The system is destined to deployed on a very large scale- cf. within the organisation vs (very) general public	Yes/No							⊗		⊗	⊗	⊗
	The system is destined to deployed on a new market	Yes/No			⊗				⊗		⊗	⊗	⊗
	The system interacts directly with the end user	Yes/No					⊗	⊗		⊗		⊗	⊗

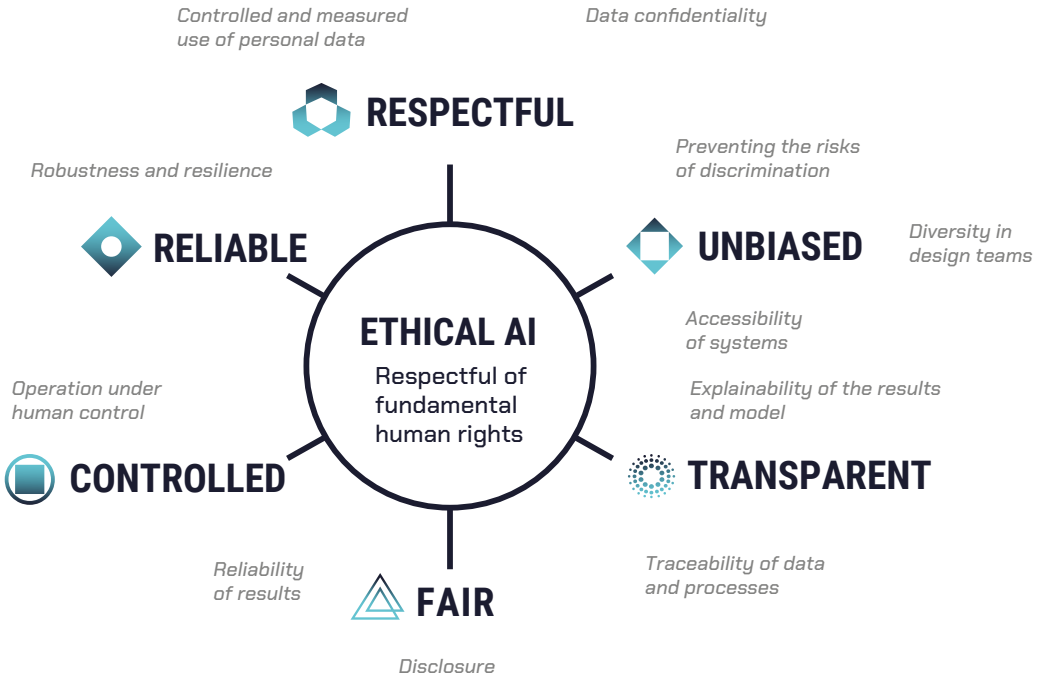
Ethical subjects to consider specifically			Data protection		No bias			Transparency		Fairness		Control	Reliability
	System purpose and implementation framework	Applies to the project	Confidentiality of personal data	Controlled and measured use of personal data	Prevention of risks of discrimination	Diversity of the project team	Accessibility of the solution	Explainability of the model and the results	Traceability of data and processes	Reliability of results	Disclosure of the AI	Operation under human control	Robustness and resilience of the solution
The technical AI solution	The system is embedded in a larger system	Yes/No							⊗		⊗		⊗
	Training the system requires a large volume of data	Yes/No											⊗
	Training the system requires the use of sensitive and/or personal data	Yes/No	⊗	⊗					⊗	⊗			⊗
	The system requires machine learning datasets from public databases	Yes/No			⊗				⊗				⊗
	The system draws on a single source to build its machine learning datasets	Yes/No											⊗
	The machine learning dataset is built from different heterogeneous databases (in terms of quality, quantity, etc.)	Yes/No			⊗				⊗				
	The system uses technologies that are by nature non-explainable (or liable to be)	Yes/No						⊗	⊗				
	The system uses «off-the-shelf» technology bricks	Yes/No											
	The system processes sensitive data - e.g. personal data, confidential data, etc.	Yes/No	⊗	⊗					⊗	⊗		⊗	⊗
	The system processes sensitive data - e.g. personal data, confidential data, etc.	Yes/No						⊗	⊗		⊗	⊗	⊗

Ethical subjects to consider specifically →			Data protection		No bias			Transparency		Fairness		Control	Reliability
↓ System purpose and implementation framework		Applies to the project	Confidentiality of personal data	Controlled and measured use of personal data	Prevention of risks of discrimination	Diversity of the project team	Accessibility of the solution	Explainability of the model and the results	Traceability of data and processes	Reliability of results	Disclosure of the AI	Operation under human control	Robustness and resilience of the solution
Project governance	The project team can refer to an in-house body in charge of ethics and AI-related subjects	Yes/No	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	The project team can refer to a set of AI project governance rules	Yes/No	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	The project team lacks diversity (gender, origin, culture, business, etc.)	Yes/No			⊗	⊗	⊗						
	The project team has been made aware of cybersecurity issues and those linked to AI in particular (cf. data poisoning, adversarial attacks, etc.)	Yes/No		⊗								⊗	⊗
	The project team has been made aware of the ethical issues	Yes/No	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	Certain actors in the system creation chain are external partners	Yes/No		⊗					⊗				⊗



GOOD PRACTICES FOR EACH PRINCIPLE

This part describes the good practices that will ensure you meet the requirements for the 6 principles. It details their implementation at every step of the life cycle of the AI system.



LIFE CYCLE OF AN AI SOLUTION

UNDERSTANDING THE NEED

- Phase when the members of the project team, the data scientists in particular, are apprised of the subject and framework of the project by their principal; this is the moment when the ethical issues are broached and the different risks, criteria and thresholds defined.

DESIGN

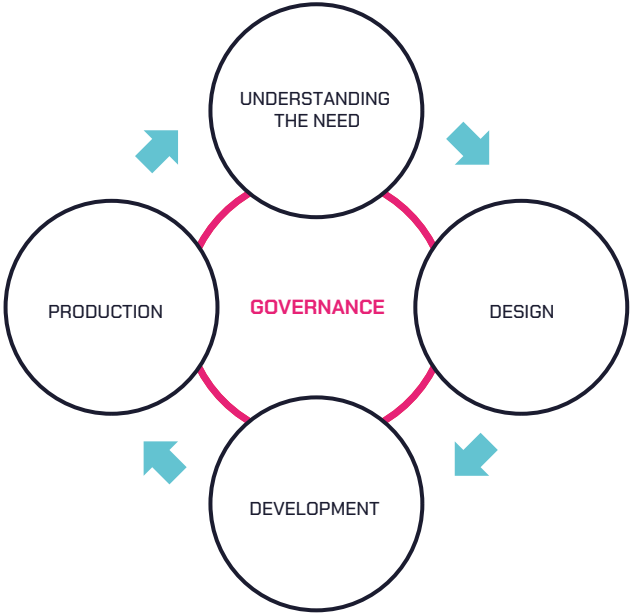
- An iterative cycle during which the data scientists build the different datasets necessary to the machine learning and testing and then prepare and test the model.

DEVELOPMENT

- The step that consists of integrating the model into its environment and developing the complementary modules that bolt onto the module itself (e.g. the interfaces); this phase generally requires computing skills.

PRODUCTION

- The AI system is operational; it is monitored and can be updated; at the end of its life, it will be decommissioned.



RESPECTFUL

«AI systems must guarantee respect for privacy and data protection throughout the entire life cycle of the system. That covers the information initially provided by the user, as well as the information generated about the user in the course of their interactions with the system (e.g. results generated by the AI system for specific users or the way users have responded to specific recommendations)»
Guidelines on ethics for trustworthy AI of GEHN IA.

EXAMPLES OF REQUIREMENTS

- The definition of what constitutes personal data varies from one region of the world to another. The same applies to the regulations on the protection of such data. the company designing an AI system must obviously comply with the regulations in force in the country where it is operating. In the European Union, the regulation to be applied is the GDPR. It rests on 5 principles that must guide the design and implementation of AI systems.
  - personal data are used for a precise, specified purpose;
  - the system only collects the personal data that are strictly necessary to it;
  - the storage time of the personal data is reasonable and decommissioning is planned;
  - there are confidentiality and security mechanisms to protect these data;
  - data subjects can access, modify and erase their data and move them from one processing operation to another.
- The system must also respect the privacy and intimacy of individuals.
- It must not use data that it has itself produced without the data subjects' knowledge.

CONTROLLED AND MEASURED USE OF PERSONAL DATA

By feeding on huge quantities of data, AI systems that operate by machine learning exacerbate the risks of abuse concerning personal data. They defy the very principles on which existing data protection regulations like the GDPR are based.

For an AI developer, compliance with these principles therefore raises a whole series of questions: what balance can be struck between strictly necessary data and performance? What does «purpose» mean in an experimental process? What path to tread between retaining data to track them and erasing them? etc.

Ideas for solutions:

UNDERSTANDING THE NEED

- Incite the actors in the project to **minimise the use of personal data** and, ideally, to do without them altogether.

- However, if personal data are necessary at any time at all during the solution's life cycle:

  - **Define the legal basis** for the collection and processing of these personal data (an obligation under the GDPR).

    → Reminder of the legal bases allowed by the GDPR (source: CNIL)
    - consent: the data subject has consented to the processing of their data;
    - contract: processing is necessary to the performance or preparation of a contract with the data subject;
    - legal obligation: processing is required by a law;

  - public interest task: the processing is necessary to the performance of a task carried out in the public interest;
  - legitimate interest: the processing is necessary for the purposes of the legitimate interests pursued by the organisation processing the data or a third party, whilst strictly respecting the rights and interests of the person whose data are processed;
  - protection of vital interests: processing is necessary in order to protect the vital interests of the data subject or a third party.

- **Carry out a Privacy Impact Assessment (PIA)** to assess the impact of the processing on these data. In certain use cases, for example if health data, the constant surveillance of persons, or personalisation and targeting tools, etc. are involved, a PIA is an obligation under the GDPR. The CNIL provides a tool for this purpose.

▶ Carry out an assessment of the privacy impact of the **loss, alteration or unauthorised disclosure** of the data handled (even where it is not personal data).

DESIGN

- **Generally, control the sources of your test and machine learning data** to avoid breaching the obligation to process data lawfully (regulatory constraints,GDPR). For example, avoid web scraping (which consists of collecting data more or less randomly without applying any vigilance).

- **Minimise and, if possible, eliminate the use of personal data** by using **synthetic data** as early as possible in the development process. This measure will be especially useful if a provider involved in the development of the AI system needs to access the data.

- If the use of personal data is unavoidable in the development and test phases:
  - collect them and process them according to the chosen legal basis (GDPR requirement);
  - make sure they are treated as confidential (see below for the different possible approaches);
  - set up access control systems (and monitor the system in production via the logging feature) to prevent any diversion of data for other purposes by other developers (GDPR requirement);

  - document the different measures taken to comply with the GDPR and in particular that justify the use of personal data in the design and development process (purpose of the application, GDPR accountability requirement).
    → See the tool Datasheets for datasets

- **Note that the retention of data** used to design the model (machine learning and tests) may be necessary (for tracing purposes and accountability). This is possible **on the condition, however,** by virtue of the principle of the data storage period in the GDPR, of:
  - justifying the storage of the data and providing information of the duration;
  - limiting the storage strictly to the length of time necessary to achieve the purpose (which implies planning for their «decommissioning»).
    → **An idea:** create multiple AI profiles corresponding to different training datasets to adapt to the user's requirements once the system is in production (this approach requires a technical and economic feasibility study).

DEVELOPMENT

- Plan for the possibility **of ending the collection** of data at any time, if the user requests it.
  → **An idea:** Stop Collect button . This would mean being ready to offer a limited or lower-performance service to the user, similar to what currently happens when you refuse the cookies in a website.
  - **Caution:** however, if the refusal of data collection leads to refusing or restricting access to a so-called essential service or a fundamental right, you will need to provide an alternative to allow the user to access the service in question.

- Develop **clear interfaces** describing the use made of personal data and allowing the user to **modify/erase** their data at any time.

PRODUCTION

- Maintain **control over the models deployed:**
  - check regularly, depending on the sensitivity of the system and the risks, that the initial purpose is still compliant and that the measures taken to respect personal data still do so (the GDPR requires that such checks be made one a year: consistency check);
  - prevent the reuse of models for other purposes or by other developers (to avoid purpose diversion (GDPR requirement), by setting up access control systems and the monitoring the system in production via the logging feature (see section Reliable/Robustness and resilience).

- Put in place a **process for alerting the data controller** if the requirements are not met.

CONFIDENTIALITY OF PERSONAL DATA

Respect for data confidentiality requires the setting up of security systems to prevent unauthorised access to databases (see section Reliable/Robustness and resilience). It will also involve using techniques preventing an individual being traced from their personal data.

There are several possible solutions. Taken individually, however, none of them is totally foolproof. Furthermore, their use will tend to reduce the system's performance.
It should also be noted that, in some cases, the possibility of re-identifying persons on whom the model has issued conclusions may prove to be necessary (e.g. AI algorithms applied to personalised medicine).

Often the best solution will be to combine several approaches, giving more weight to one or other of them according to the purpose of the application and the performance objectives.

Ideas for solutions:

DESIGN

- Apply a combination of methods to keep datasets confidential.
  - **Anonymisation:** there are several techniques to choose from, and the choice will depend on the purpose of the use case and the characteristics of the datasets. If the basic techniques are not enough (e.g. if they make it possible to re-identify data subjects by cross-checking with information in other databases), more sophisticated methods can be tried, such as differential privacy, which consists of adding noise around the data points collected in order to «drown them out». Another technique, known as Avatar anonymisation, which is used by the WeData company, has recently been approved by the CNIL.
  - **Pseudonymisation:** these are methods that consist of replacing directly identifying data in a dataset (surname, first name, etc.) with «indirectly identifying» data (aliases, sequential numbers, etc.). It is also possible to envisage collecting less precise data (age group rather than exact age, postcode instead of a full address, etc.).

- **Distributed and/or federated learning:** these approaches aim to reduce the centralisation of data, to keep them as close to the entity that generates them as possible and not to expose them.
  In distributed learning, the algorithm accesses a database that remains located on its owner's premises. In federated learning, the algorithm accesses a distributed network of databases (collaborative operation).
  → See the approach recommended by the Substra Foundation.
- **Data encryption:** this is a common approach in projects where confidential data are pooled. An example would be the health data on the Health Data Hub, which are encrypted.
  → For more detail on this subject, see the methods described by Substra Foundation.

- Also protect the confidentiality of the models (which could by inference reveal the data used for the machine learning). For example, by carrying out the machine learning by knowledge distillation (which has the added advantage of compressing the model).
- Measure the effectiveness of anonymisation, by analysing the risk of re-identification.
  → For example using tools like ARX: Risk analysis - Data Anonymization Tool.

- Protect datasets transferred to partners using the methods described above.

- Document any vulnerabilities and the techniques used to overcome them.

EXPLORE THE SUBJECT FURTHER

- Audit requirements for personnal data processing activities involving AI : le guide méthodologique de l'autorité espagnole de protection des données pour évaluer la conformité d'un système d'IA au RGPD (2021).
- Guide RGPD du développeur de la CNIL.
- De la difficulté technique de l'anonymisation ou comment mal anonymiser ses données , un article publié sur Medium par Wavestone (2018).
- Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle que l'on peut trouver sur le site Éthique et intelligence artificielle de la CNIL (2017).

UNBIASED

«The datasets used by AI systems (for both training and operation) can be distorted by accidental historical bias, omissions and defective governance models. The persistence of such biases could be a source of involuntary (indirect) discrimination and prejudice.»
Guidelines on ethics for trustworthy AI of GEHN IA.

EXAMPLES OF REQUIREMENTS

- The system must operate impartially. In particular, it must aim not to reinforce or create discrimination due to biases introduced in the training process or in the algorithm.
- It must reflect the diversity of the user or other population concerne
- It must comply with the most common standards of accessibility and foster diversity and inclusiveness.

PREVENTING THE RISKS OF DISCRIMINATION

A machine learning AI system can include biases (e.g. if it has been trained with a dataset that was itself biased). The risk in this case is of producing false or discriminatory and therefore prejudicial results. Hence the requirement to strive to achieve unbiased systems. Nevertheless, attempting to find a complete and final solution to this issue amounts to trying to square a circle. For a start, the notion is eminently culture-bound. What's more, it comes in different forms, which are sometimes incompatible with each other. Depending on the use case, it will therefore be necessary to establish «non-bias» criteria, whilst ensuring they are compatible with the ethical framework laid down by the company, and to attempt to measure the risks of bias and then try to get as close as to an optimum situation as possible.

Ideas for solutions:

UNDERSTANDING THE NEED

- If the conclusions of the AI system concern people:
  - carry out **an analysis of the risks of discrimination and an impact assessment** (see section on Risk assessment),
  - and, if a risk appears, examine the parameters and variables that could directly or indirectly generate a risk of discriminatory bias or cause drift in in the model initially defined during its use.

- Where possible, contact:
  - **specialists in the human scientists** (sociologists, anthropologists, etc.), who can contribute their expertise and help you to identify the risks of potential biases,
  - and/or **statisticians** to work on the structure of the machine learning datasets and circumvent the statistical pitfalls.

DESIGN

▶ **Building datasets.**
→ **An idea:** make non-bias a criterion of success (on a par with performance) in order to orient the work and keep the issue in the top of your mind as you go forward.

- Always make sure you are **in control of the datasets** used for machine learning and testing: ask yourself where they come from and how they were built (know the source, the distribution of the dataset, how the data were collected, what transformations they have undergone, etc.).
- Several construction possibilities:
  - create high-quality datasets, calling upon experts in the field concerned;
  - use documented existing training datasets;
  - if the datasets are not representative or extensive enough, use the different statistical techniques available to make up the deficiency (augmentation, re-sampling, etc.) or test the use of
  - generative adversarial networks (GANs) to generate synthetic data;
  - envisage pooling your data with other companies in the same field, at national or European level (see the Voice data example).

▶ **Designing the model.**

- Integrate these constraints into the algorithm.

▶ **Measuring and correcting.**

- Instrument the test process to facilitate its implementation.
- Validate the representativeness of the dataset, e.g. using existing reference standards (institutional, open data, etc.).
- Test the algorithm against the non-bias criteria initially defined.
  → An approach defended by the Institut Montaigne in its white paper Algorithmes, contrôle des biais SVP, which it calls «active fairness», consists of using datasets containing protected variables (the approach requires submitting an impact assessment to the CNIL).
- Make sure you have not introduced any new biases while trying to correct the initial ones.
- Adjust the machine learning algorithm, if necessary.
  → **An idea:** build bias-correcting algorithms (processing bias by bias) using GANs.

PRODUCTION

▶ Plan to continuously monitor for drift, using automated supervision systems and appropriate metrics (thresholds, etc.).

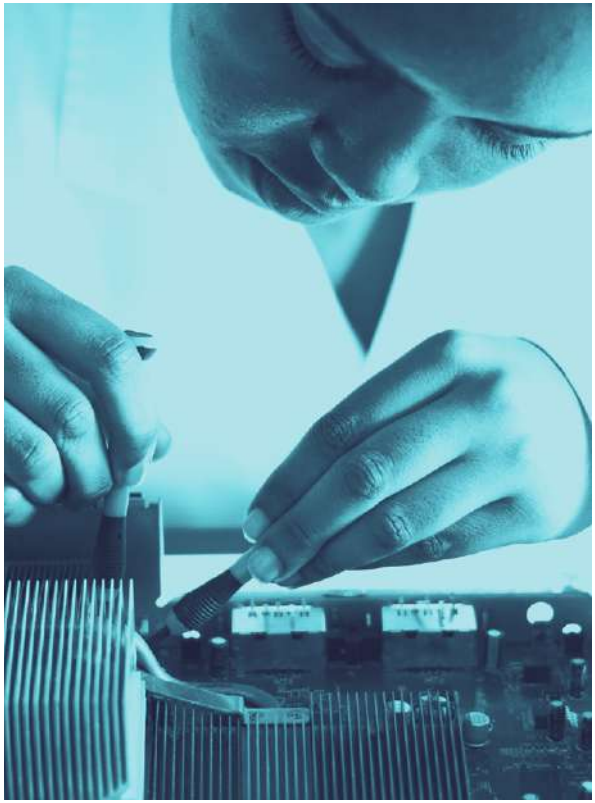
▶ Appoint a single point of contact so that users can submit their observations.

DIVERSITY IN DESIGN TEAMS

The aim of the diversity requirement is to help project teams to consider the risks of bias in their products under the best conditions and keeping as open a mind as possible. It also tends to encourage teams to embody their ethical values in their products.

Ideas for solutions:

- Foster diversity in your teams (gender, culture, background, etc.).
- Work with and recruit from inclusive educational institutions.



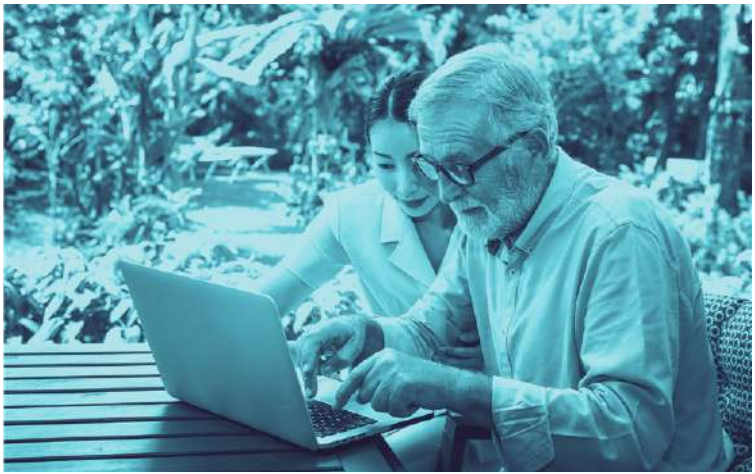
ACCESSIBILITY OF SYSTEMS DES SYSTÈMES

AI has real potential for facilitating access to the digital world for the disabled or those who have experienced difficulties with digital technology. Beyond the regulatory accessibility functions a graphical user interface must have, equipping AI systems that interact with humans with image recognition, text-to-speech or text analysis technologies, for example, would also facilitate the accessibility and inclusiveness of these tools.

Pistes de solutions :

DESIGN

- Comply with **regulatory obligations** on graphical user interface accessibility.
- Use the Référentiel Général d'Amélioration de l'Accessibilité (RGAA), a French standard on the improvement of accessibility.



EXPLORE THE SUBJECT FURTHER

- Algorithmes, biais, discrimination et équité, a white paper produced by Patrice Bertail, David Bounie, Stephan Clémençon and Patrick Waelbroeck at Télécoms ParisTech (February 2019).
- Igorithmes, contrôle des biais SVP, a white paper published by the Institut Montaigne (March 2020).
- Unfair biases in Machine Learning: what, why, where and how to obliterate them, an article on the origin of discriminatory biases in machine learning algorithms and how to overcome them, by Paul Irolla (MS Security, 2020).
- A tutorial on fairness in machine learning, atechnical post by Ziyuan Zhong (Towards Data Science, 2018).

TRANSPARENT

«Individuals must be able to understand how AI systems make decisions, especially when they impact their daily lives.»
Notre approche IA responsable, IA fiable, Microsoft.

EXAMPLES OF REQUIREMENTS

- The causes and criteria that lead to AI conclusions must be able to be made known to users and/or the people affected or concerned by the conclusion; these explanations must be intelligible in order to shed light on the final decision or act on input data.
- The process of designing/developing/deploying the system must be documented (description of data collection and labelling, the algorithm used, the business model, etc.) for the purposes of verification (audit) and improvement/correction of the tool.
- The process of collecting, storing and using the data must also be documented (in line with the GDPR).
- The different trade-offs that have an effect on the ethical requirements must be justified.

EXPLAINABILITY OF RESULTS

An AI system's ability to **make the impact of a variable on a result explicit** and, more generally, the **comprehension of the processing** carried out by a AI system, are decisive factors in the acceptance of AI by society. It is sometimes also a compliance and/or security requirement (the CNIL requires that information be provided at least on the data used to arrive at a result; the GDPR requires an explanation on whether personal data are involved in the result). Unfortunately, the AI systems that currently perform at the highest level turn out, as soon as substantial quantities of data begin to be processed, to be the most opaque... And yet there are solutions to resolve this question.
They can be implemented until the many research projects underway in this area bear fruit.

Pistes de solutions :

UNDERSTANDING THE NEED

- Estimate the need for and degree of explainability required, **depending on the use case and the purpose of the product.** The problem will be more or less acute according to the use case. For example, it is likely that we will give less importance to the explainability of a book recommendation service in an e-commerce site than to a rating algorithm for granting bank loans. And yet, it is quite possible that an e-commerce service might want to set itself apart by backing up its recommendations with an explanation based on the ethical goal of combating intellectual narrow-mindedness.

- If there is clearly a need for explainability, work with relevant experts to define the **trade-offs to be made** between accuracy of results and transparency. In many cases, at least at the outset, the use of a naturally explainable and less precise algorithm will be enough to meet the need.

- Envisage hybrid systems, which can embed more or less opaque machine learning algorithms in algorithms based on perfectly interpretable rules.

- The question of explainability raised at regular intervals throughout the design and development process as the results improve.

DESIGN

- If it is necessary to use a complex algorithm that is not naturally explainable, look to **post-hoc explainability methods** (e.g. Lime, Shap). The choice of approach will depend on the **use case** and **the target of the explanation.**
  - A customer, consumer or end user will want an explanation based on precise variables (their own or those corresponding to a specific context). The system must provide a local explanation (e.g. Lime or Anchor).
  - A professional will want to understand the model as a whole and will expect a global explanation with a view to improving it (e.g. Shap).
  - The regulator or supervisor will want proof and will need a global explanation (e.g. Shap).

- **Test out** the different explainability approaches to draw common conclusions.

- **Document** the different approaches tested and record the different results to draw common conclusions.

PRODUCTION

- Use **graphic tools** that allow you to visualise the dominant criteria.
  → A tool like Shapash, designed by the MAIF insurance company, makes the results provided by the most common explainability tools (Lime, Shap, etc.) accessible to non-specialists.

TRACEABILITY OF PROCESSES AND DATA

Tracking the datasets and the design processes and methods used on the different versions of the model is a prerequisite for being able to check the absence of bias, non-diversion of purpose and the reliability of the systems.
This means documenting everything connected to an AI system. It should be noted that, in fact, this amounts to nothing more and nothing less than applying a quality approach, as is common practice in other areas of IT.

Pistes de solutions :

DESIGN AND DEVELOPMENT

- **Document:**
  - the machine learning phase: data used (sources, transformation), parameters and hyperparameters, algorithm, versions, etc.:
    - for this, you could take your inspiration from the Google model card, which is a sort of ID card for a model,
    - or you could use the Datasheets for datasets method suggested by the team of researchers that developed it;
  - the different trade-offs between performance, explainability, confidentiality and security;
  - the people involved in constructing the AI system and their roles.

- Set up a reference base where you will store all your AI-related information.
  → **An idea:** create a «family tree» of your system along the lines of the model being developed by the Substra Foundation.

PRODUCTION

- Continue documenting the data and behaviours of the AI system once it is in operation:
  - set up a **logging system** to record the contexts in which results are obtained:
    the algorithm, the version, the model, the parameters and hyperparameters, the dataset (bearing in mind that this system can quickly get extremely complex if the system is continually learning);
    → **Example of a versioning tool:** DVC.org (Open-source Version Control System for Machine Learning Projects)

plan the **conditions of storage** of this information (logs can contain personal or sensitive data*): storage time and purpose to be specified, security measures to be taken, access rights to be granted, etc. (see Respectful section).

**It should be noted, that in certain cases, the strict application of the GDPR comes up against some serious operational difficulties: for example, if the application is a cybersecurity system that records thousands of IP addresses a minute, which may be considered as personal data by the CNIL.*

EXPLORE THE SUBJECT FURTHER

- Interpretable Machine Learning or how to make black box models explainable: a regularly updated guide by Christoph Molnar.
- Datasheets for datasets : a guide to tracking your datasets produced by a team of researchers from Google (actually, Timnit Gebru), Microsoft and different universities (2020).

FAIR

«Each person must be aware when he or she is interacting with a machine.»
Rome Call for AI Ethics.

EXAMPLES OF REQUIREMENTS

- The user must unambiguously understand that they are interacting with a machine.
- The area of intervention as well as the limits and capacities of the system must be made known to the person that is going to use it.
- The user must be aware whether an AI system is involved in the result of a calculation which could be decisive for them or orient a decision concerning them or that they have to make.
- The system must do what is expected of it, no more, no less.

DISCLOSURE

The system's ability to disclose itself, that is to say reveal what it is and what it does, is a key factor in establishing users' trust and reducing the risks of abuse of ignorance and addiction. This issue concerns AI systems that interact directly (e.g. chatbots) or indirectly (if the system is embedded in a tool) with humans. Nonetheless, it should be noted that a system that says too much will be more exposed to inference attacks.

Ideas for solutions:

UNDERSTANDING THE NEED

- Ensure that all the stakeholders in the project are informed of the issues linked to the disclosure of the system so that they can provide the information be communicated and the areas and channels of communication with the user can be defined.

DESIGN

- Set up a system or method to inform the user that they are dealing with an AI system.
  → **Example:** in the case of a chatbot, inform the human from the outset that they are interacting with a robot or design an interface that is sufficiently explicit to avoid any ambiguity (e.g. avoiding using a human avatar)

- Generally, provide a summary that is easily read and understood by the user, informing them of:
  - the fact that an AI system is involved in the solution used;
  - the type of AI at work, its area of intervention and its objectives;
  - its limits and the potential margin of error;
  - the conditions in which it must be used and any risks involved in using it;
  - and, where applicable the use made by the system of the personal data asked for (GDPR purpose requirements; see section Controlled and measured use of personal data).

RELIABILITY OF RESULTS

This subject goes hand-in-hand with disclosure, where the system reveals what it is and what it does. Here the system is deemed fair and reliable if it does what it says. This issue is one of the most complicated to deal with, as by nature, AI systems operating by machine learning cannot guarantee 100% reproducibility of their result, and even less so if they are continually learning. For complex systems handling a very large number of parameters, one of the problems that arises is the impossibility of being able to test and therefore validate all the possible cases.

Ideas for solutions:

UNDERSTANDING THE NEED

- Establish binding rules and limits not to be exceeded with regard to the results of the risk and impact assessments carried out earlier.

DESIGN

- Write a detailed set of specifications for the system and specify what results are expected so as to provide at least partial proof of the reliability of the model.

- Control the code:
  - make compromises between simplicity and performance to have better control over the reliability of the results;
  - prefer documented, open source and academic codes (bearing in mind, though, that this brings with it a risk of becoming a target of a cyber attackers who can also access the code).
    → **An idea:** pair AI systems with each other with voting methods to smooth out the errors.

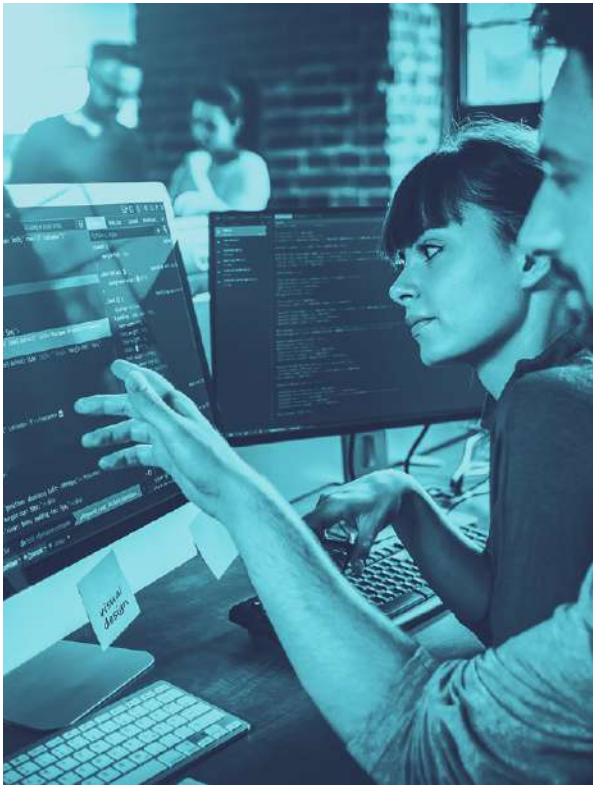
- Control the training and test datasets:
  - use documented training datasets;
  - audit the sets and, in particular, assess the representativeness of the data.
- Tracking and tracing (see Traceability section):
  - document everything: sources of data, processing of data, model, algorithm, training methods;
  - create a CI of the model, which will develop over time.

DEVELOPMENT

- Have the results of the AI systems checked by business experts to guarantee the reliability of the results.
- Automate tests to standardise them and make them easier to perform.
- Test all the possibilities of the system if it is possible
  - Whether or not it is possible to implement this measure will, of course, depend on the complexity of the use case. Envisage using GANs to generate test data, if necessary.

PRODUCTION

- Establish test and behaviour checking protocols throughout the entire service life of the system to ensure there is no drift.
- Check the viability of the tests.



EXPLORE THE SUBJECT FURTHER

- Reproducible operations – commitment to the #4 principle, Pages from the website of the British Institute for Ethical AI & Machine Learning on the reliability of machine learning systems (the site is very comprehensive and covers all the ethical issues around machine learning).
- Guidelines for the development of responsible conversational AI by Microsoft.

CONTROLLED

« Only human beings can be held responsible for decisions stemming from recommendations made by AIS, and the actions that proceed therefrom.»
Montréal Declaration.

« In all areas where a decision that affects a person's life, quality of life, or reputation must be made, where time and circumstance permit, the final decision must be taken by a human being and that decision should be free and informed.»
Montréal Declaration.

EXAMPLES OF REQUIREMENTS

- The user of an AI system is made aware of the recommendations made, but must still be able to make their own autonomous decisions and personal choices.
- If the AI system's conclusion will lead to a decision that affects one or more persons, then the final decision must rest with one person.
- The human must be able decide not to use the AI if they believe the ethical or security conditions are not met.
- The system must allow a person to contest it or report an anomaly.

OPERATION UNDER HUMAN CONTROL

Risk of dehumanisation of our societies and loss of the decision-making autonomy of the individual, issues around accountability (a human must remain solely accountable for the actions and decisions of an AI system). The reasons behind these requirements are many and varied. From providing a «contact» button to restricting the amount of automation in the system, the measures to be taken will depend on the use case.

Ideas for solutions:

DESIGN

- If the results have consequences for humans, keep the automation of decision-making to a minimum. The system must remain an aid to decision-making.

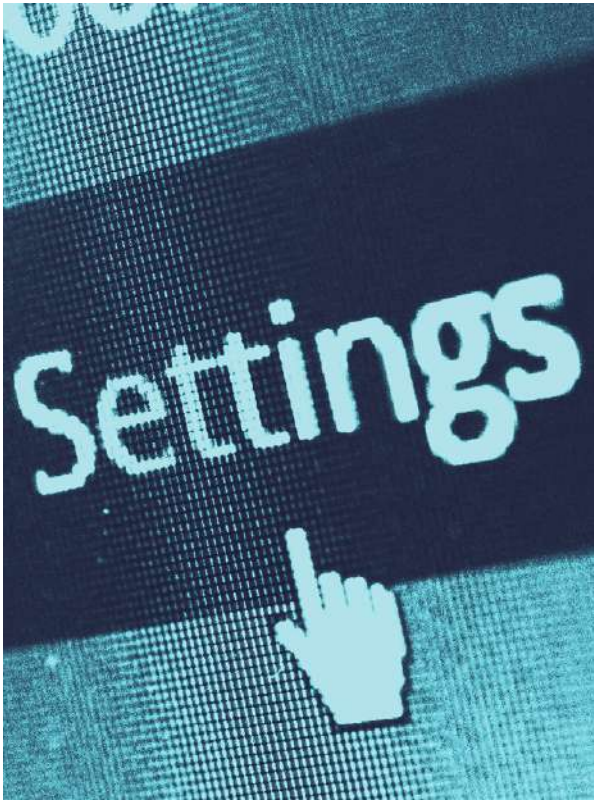
DEVELOPMENT

- For systems aimed at the general public, provide for the possibility of adjusting the parameters and data in the model that concern the user in an interactive way and in real time, or set up a system enabling the user to validate their parameters at the time of use.
- In certain contexts involving the general public, in particular in the field of public services, inform the user that AI is part of the system (the disclosure requirement) and provide for the possibility of the user choosing not to use the AI element [RGPD requirements/Art 22].

PRODUCTION

- Create a «Contact» button allowing the user to interact, to transfer information or lodge an appeal, and set up an internal process allowing users' appeals and feedback to be dealt with.

  → **An idea:** show the user all the AI scores and not just the «final» result. This approach supposes that the user will learn about the functioning of the AI system.



EXPLORE THE SUBJECT FURTHER

- Fully automated decision making AI systems: the right to human intervention and other safeguard, a methodological guide that can be consulted on the Ai Auditing framework website of the British data protection authority.

«AI systems and the environments they operate in must be safe and secure. They must be technically robust and care must be taken to ensure they are not exposed to malicious use. »
Guidelines for trustworthy AI of GEHN IA.

RELIABLE

EXAMPLE OF A REQUIREMENT

- The system must include mechanisms allowing it to:
  - protect itself against the risks of wrongdoing identified in the risk and impact assessments;
  - protect itself against cyber attacks (purpose diversion, data theft);
  - block and correct the effects of any attacks and malicious actions.

ROBUSTNESS AND RESILIENCE

Like any computer system, AI systems are concerned by cybercrime. They must therefore be protected by the same rules, principles and systems as any other information system. But they are also subject to certain specific cyber threats:

- **Data poisoning:** the attacker seeks to skew the behaviour of a model by modifying the learning data. Continually learning systems are particularly exposed to this type of attack.
- **Evasion:** here the attacker tampers imperceptibly with the application's inputs to deceive the system and induce a decision different to the one normally expected (e.g. the panda). Systems processing complex input data such as images are particularly sensitive to this type of attack.
- **Inference:** the attacker bombards the AI system with queries to understand how it works and grasp its key parameters with the aim of imitating the system. Systems that disseminate a lot of information are easily exposed to this type of attack.

How can you protect against these threats and guarantee operation of the system without damaging the integrity of persons or breaching any of the ethical values?
The first measure to take, irrespective of the projects themselves, consists of raising the data scientists and data analysts' awareness of cybersecurity issues. Unlike computer scientists, this population, who generally have a background in mathematics or statistics, are naturally less preoccupied with such matters (see Governance section).

Ideas for solutions:

DESIGN

- **Secure the machine learning process** to reduce its exposure to data poisoning attacks:
  - control the sources of machine learning data;
  - protect access with access control and accreditation systems;
  - reduce the quantity of data to a minimum, and especially sensitive data necessary to machine learning (use synthetic data, if possible);

- apply techniques to reinforce the confidentiality of sensitive data (anonymisation techniques using differential privacy, pseudonymisation and distributed and/or federated learning (see Data confidentiality section);
- continually monitor the progress of the learning and any changes in the behaviour of the models;
- put safeguards in place, e.g. for a chatbot, create a blacklist of terms to block, in both input and output;
- apply practices like RONI (Reject on Negative Impact), which consist of rejecting any data that causes drift in the model.

► **Reinforce the robustness of the models** to reduce their sensitivity to different types of attack, in particular inference and evasion attacks.
- Different techniques exist: knowledge distillation, adversarial learning, noise addition on input data, etc. These will be applied according to the level of precision you want to achieve.
- Reduce to a minimum information that could reveal how the model works (score, precision, etc.) to reduce inference attacks. This measure will be applied according to the level of disclosure deemed indispensable.

DEVELOPMENT

► **Test:** set up (automated) processes and tools such as IBM's ART 360 toolkit or the code alteration approach.

PRODUCTION

► **Monitor:** set up a process of real-time log monitoring. In particular check that there is no drift in the results, by regularly having business experts check the behaviours and results.

► Create a contact button so that the user can alert the data controller.
→ **An idea:** create a Bug Bounty platform.

EXPLORE THE SUBJECT FURTHER

- Artificial Intelligence and Cybersecurity, a white paper from Wavestone (2019).

PART 3

USE CASES AND EXAMPLES

YOUR NOTES:

E-COMMERCE USE CASE

DESIGNING AND IMPLEMENTING A RECOMMENDATION ALGORITHM FOR SITE SELLING CULTURAL GOODS ONLINE.

A potential customer (prospect) goes onto an e-commerce site selling cultural goods. They browse the catalogue of products on offer to find what they are looking for. The site recommends products while they are browsing.

ETHICAL RISKS IDENTIFIED

Examples of specific issues the e-commerce company should address:

- **discrimination** according to the customer's gender/religion/living standard, etc.
- **being enclosed** in a personal bubble of personalised content;
- **being influenced** by the interface;
- **use of the customer's unconscious biases** to influence their purchases (nudge).

Description of the use case

Aims (regarding the user)	To enrich the customer's experience by personalising the content presented to thWem to meet their needs, without distorting the selection submitted to them in such a way as to limit their choice.
Purposes (for the company using the AI system)	To increase its sales (maximise the conversion rate, optimise the average basket, etc.).
How	By turning the prospect into a customer by achieving an effective purchase. To do this, it is therefore necessary to offer the product that corresponds best to what they are actually looking for in terms of the product itself, price, access (collection/delivery), etc.
Actions taken	Optimisation of the presentation of the product - level of product description, photos, etc. Presentation of similar and/or complementary products.
Solution adopted (type of AI)	Model/Algorithm recommended.
Methods/Tools used	Collaborative filtering (user-based/item-based) Content-based (proximity by product type). Popular (association model - cf. association of products in the buying act).
Who is concerned?	At the Understanding the need stage: product marketing, distributors, product producers, consumers. At the design stage: data scientists, data engineer, designer. At the development stage: developer, testers. At the deployment stage: supervisor, testers.

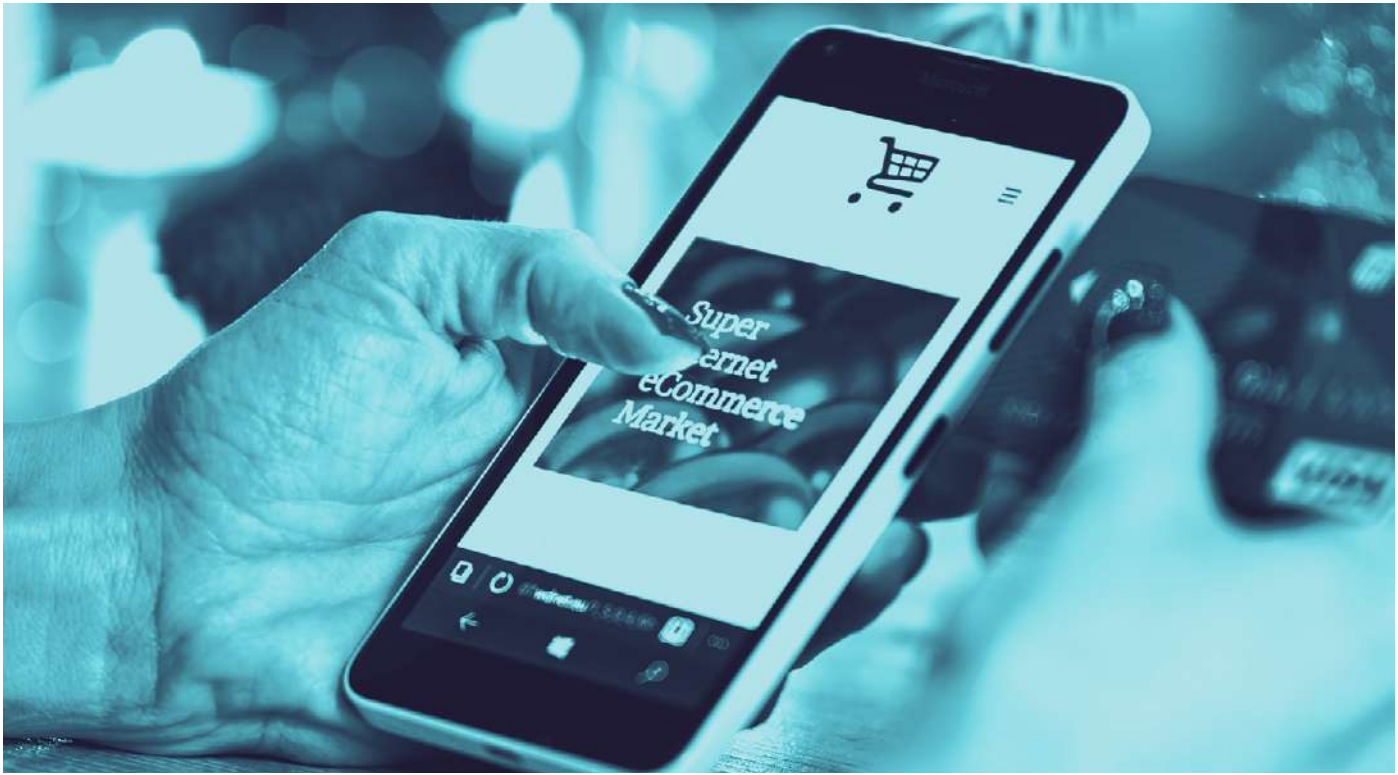
Assessment of the solution's sensitivity to ethical issues (based on the ethical sensitivity matrix)

SYSTEM PURPOSE AND IMPLEMENTATION FRAMEWORK		Applicable to the project
The business need	The system automates a decision, or helps to make a decision concerning physical persons	Yes
	The system automates the performance of tasks for the user	No
	The system is destined to deployed on a very large scale - cf. within the organisation vs (very) general public	Yes
	The system is destined to deployed on a new market	No
	The system interacts directly with the end user	Yes
The technical AI solution	The system is embedded in a larger system	Yes
	Training the system requires a large volume of data	Yes
	Training the system requires the use of sensitive and/or personal data	No
	The system requires machine learning datasets from public databases	No
	The system draws on a single source to build its machine learning datasets	No
	The machine learning dataset is built from different heterogeneous databases (in terms of quality, quantity, etc.)	Yes
	The system uses technologies that are by nature non-explainable (or liable to be)	No
	The system uses "off-the-shelf" technology bricks	Yes
	The system processes sensitive data - e.g. personal data, confidential data, etc.	No
	The system is constantly learning	Yes
Governance of the project	The project team can refer to an in-house body in charge of ethics and AI-related subjects	Yes
	The project team can refer to a set of AI project governance rules	Yes
	The project team lacks diversity (gender, origin, culture, business, etc.)	Yes
	The project team has been made aware of cybersecurity issues and those linked to AI in particular (cf. data poisoning, adversarial attacks, etc.)	No
	The project team has been made aware of the ethical issues	Yes
	Certain actors in the system creation chain are external partners	Yes

Subjects to address
(according to the ethical sensitivity matrix)

Ideas for solutions identified
(for each subject to be addressed)

ETHICAL SUBJECTS TO CONSIDER SPECIFICALLY		ACTION(S)/PHASE(S) OF THE PROJECT/ACTOR(S)
Respect	Confidentiality of personal data	**Design:** pseudonymisation of personal data, setting up of secure access to data, list of features that are inputs for the algorithms, work from customer identifier, non-integration of data considered as «sensitive» within the meaning of the GDPR
	Controlled and measured use of personal data	**Design:** two separate data processing operations: learning and use of the model
Unbiased	Prevention of risks of discrimination	**Development:** validation of data relevant to operation, final minimisation
	Diversity of the project team	**Design:** introduction of random data for deliberate diversity in the recommendation **Development:** diversity in the test team or test roles
	Accessibility of the solution	**Design:** design for accessibility to all audiences from the outset
Transparency	Explainability of the model and the results	**Design:** need for transparency for the tester and the customer **Deployment:** propose different parameter settings to the user: relevance, dates, etc.
	Traceability of data and processes	**Design and development:** creation of a data catalogue, project documentation (CDM, data input process, etc.) to guarantee traceability in a governance tool
Fairness	Reliability of results	**Design:** definition of the objective and metrics **Development:** measurement and initiation of monitoring **Deployment:** performance monitoring, transfer of alerts
	Disclosure of the AI	**Design and development:** provide for an alert message for the end user
Control	Operation under human control	**Deployment:** supervision chain, monitoring, user feedback loop via customer relationship
Reliability	Robustness and resilience of the solution	**Design:** define the metrics **Deployment:** set up the monitoring and alert handling processes, regular backtesting of the models


Super
ernet
eCommerce
Market

EXAMPLE OF A PROJECT

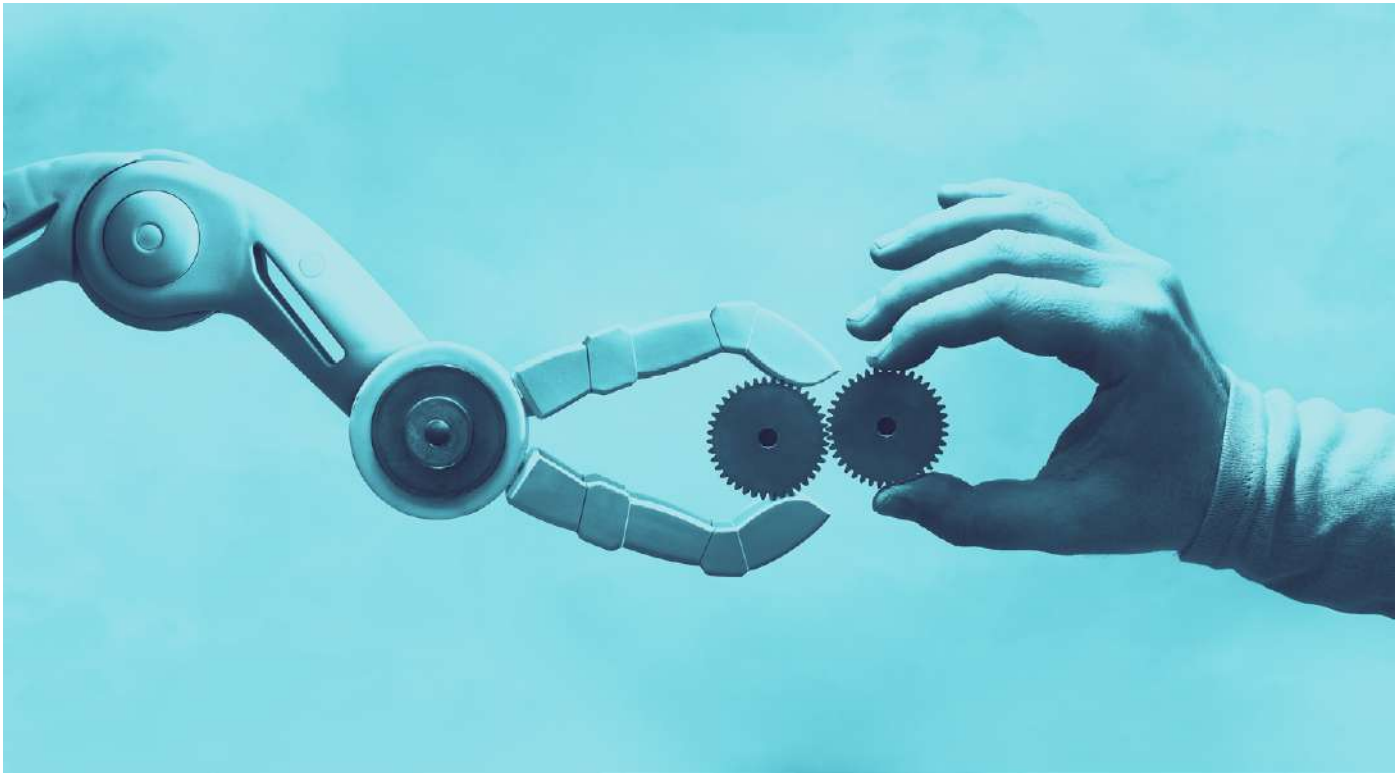
IMPLEMENTATION OF A TRACEABILITY SYSTEM

- Certain artificial intelligence development scenarios require particular precautions in use. A typical example of such a case would be critical projects where AI operates in real time and without human intervention, leading to a risk of damage to the customer's reputation or material damage or even injury to persons, if the system were drift off course. Another scenario that requires particular precautions arises when the team in charge of integrating the model is not the same as the R&D team that created it, as this can potentially lead to communication problems: the risk is higher, in particular, in the Open Source world or in large-scale projects comprising separate operational units.

- To deal with these issues of strict control concerning the learning and use of artificial intelligence models, we also use a standardised organisation in the form of the «model cards» developed by the research teams at Google in 2019, whose aim is to document an AI model succinctly, but comprehensively. A model card takes the form of a summary in about twenty lines, a bit like the Readmes that come with some software, containing the following information:
  - details of the model, and in particular: identity of the developers, publication date, parameters of the model, version, licence;
  - use cases planned and nominal conditions of use;
  - characteristics of the datasets used for machine learning;
  - characteristics of the model and datasets used for the assessment,
  - ethical considerations, warnings and recommendations.

- Adopting this standard means that you have a common template for the documentation of each artificial intelligence system, thereby allowing the transmission of a maximum of information between the engineers and developers to ensure that the models are used correctly. And as a result, this reduces the risk of unforeseen behaviour by the associated automated systems. .

Mathis HAMMEL
Head of Cybersecurity R&D Sogeti (Groupe Capgemini)



Numeum and its partners warmly thank all those who took part in the work.

► STEERING COMMITTEE

- Céline BAYLE (SAGE)

- Bénédicte DE LINARES (BDL CONSEIL)

- Valentin HUEBER (NUMEUM)

- Katya LAINÉ (TALKR.AI BY KWALYS)

- Jean-Claude REMBERT-BAUDET (ASTEK)

► MENTORS AND COORDINATORS

- Magali BARNOIN (TELECOM VALLEY)
- Benoît BOUFFARD (WAVESTONE)
- Marine BROGLI (DPO CONSULTING)
- David CORTES (AI-VIDENCE)
- Laurence DEVILLERS (SORBONNE UNIVERSITÉ/CNRS)
- Sébastien JARDIN (IBM FRANCE)
- Mouchira LABIDI (FREELANCE)
- Charlotte LISCHER (CATALIX)
- Alice LOUIS (CABINET DICÉ)
- Jean-Luc MAINGUY (SEENAPSYS)
- Fabrice MARQUE (ZEBRAVALLEY)
- Emmanuel NARS (DOCAPOSTE)
- Vincent PERRIN (IBM FRANCE)
- Françoise SOULIÉ (HUB FRANCE IA)
- Florence TRESSOLS (IBM FRANCE)
- Félicien VALLET (CNIL)

► POUVOIRS PUBLICS

- Renaud VEDEL (CSN-IA)
- Nicolas AMAR (CSN-IA)
- Martin BIERI (CNIL)

► CONTRIBUTORS

- Sonia ABECASSIS (IBM FRANCE)
- Cindy ACCOLAS (GRAND ENOV +)
- Didier AÏT (OPTIM'EASE)
- Marianne ALLANIC (ALTHENAS)
- Aziz AMAL (ASTEK)
- Nadia ANGLESSY (NETSYSTEM SOLUTIONS)
- Nicolas Andréa ARZOTTO (LEADIN)
- Marion BALAC (ESAM)
- Franck BARDOL (DIAG)
- Renaud BAUVIN (CRITEO)
- Julie BEC (AIR FRANCE KLM GROUP)
- Jérôme BERANGER (ADELIAA)

- Gwenaëlle BODILIS (DPO SYSTEM)
- Marina BOECHAT (MYDATAMODELS)
- Eric BONIFACE (SUBSTRA FOUNDATION)
- Guillaume BUFFET (U CHANGE)
- Anne-Christine CARPENTIER (GFII)
- Pierre CHARARA (TESSI)
- Lucas CHARRON (SPORTINTECH)
- Edouard CHOPLAIN (C2IP)
- Tawhid CHTIOUI (AIVANCITY)
- Eugénie CLÉMENT (OCCITANIE DATA)
- Sophie COMPAGNON (CRITÉO)
- Jean-Baptiste CONAN (KEYRUS)
- Nathalie COSTA (YSANCE)
- Rébecca DADI (DPO CONSULTING)
- Guillaume DE LA ROCHE (RENAULT)
- Nathalie DELBECQ (RENAULT)
- Paul DESIGAUD (WAVESTONE)
- Alix FAUQUES DE JONQUIERES (ANITI)
- Sébastien FORET (GRAND ENOV +)
- Mickaël GADOUD (WAVESTONE)
- Mithuran GAJENDRAN (WAVESTONE)
- Nicolas GEORGEAULT (ASI)
- Guillaume GIMONNET (WAVESTONE)
- Emmanuel GOFFI (INSTITUT SAPIENS)
- Amélie HELIOU (CRITEO)
- Laëtitia KAMENI (ACCENTURE)
- François KLIEBER (BOUYGUES CONSTRUCTION)
- Djémila KOHIL (LPCE BIOBANK CÔTE D'AZUR)
- Bradreddine LADJEMI (ANKABOOT)
- Pascal LAINÉ (TALKR.AI BY KWALYS)
- Yanelle LARIBI (IMPACT AI)
- Yann LE BIANNIC (SAP)
- Fabrice LE GUEL (RITM)
- Frédéric LEBLAN (3DS OUTSCALE)
- Xavier LECLERC (DPMS)
- Bertrand LEJEUNE (CAP DIGITAL)
- Simon LEROY (KEYRUS)
- Clément LOMBARD (WAVESTONE)
- Daphné MARNAT (TWISTING)
- Laura MARTI (BOUYGUES CONSTRUCTION)
- Maud MARQUIS (MIO&CO)
- Didier MASCARELLI (KADLOG)
- Clément MAYER (SUBSTRA FOUNDATION)
- Igor MEKHOV (CONSORT NT)
- Stéphan MIR (WAVESTONE)
- Assia MOULOUDI (SAP)
- Claire NICODEME (SNCF)
- Bernard OURGHANLIAN (MICROSOFT)

- Pierre PARREND (EPITA)
- Alexandre PASCAULT (ASTEK)
- Stéphane PAULIN-HENRIKSSON (CNRS)
- Gaëlle PICARD-ABEZIS (DOCAPOSTE)
- Estelle PINCHEZON (HUMAN DESIGN GROUP)
- Gaëlle PINSON (HUB FRANCE IA)
- Marc PLATINI (GRAND ENOV +)
- Timothée RAYMOND (LINEDATA)
- Bernardo RESENDE (THALES SERVICES SAS)
- Bettina REVEYRON (IMPACT AI)
- Caroline RICHARD (NATIXIS)
- Laurent RISSER (ANITI)
- Céline RODAP (ECOLE 42)
- Roxana RUGINA (IMPACT AI)
- Laura SARRIOT (KILOUTOU)
- Céline SAVOY-LAMOTTE (TESSI)
- Anthéa SERAFIN (OCCITANIE DATA)
- Emilie SIRVENT-HIEN (ORANGE)
- Camille SOUILLART (HUB FRANCE IA)
- Thomas SOUVERAIN (DREAMQUARK)
- Alexis STEINER (GRAND ENOV +)
- Aurélia SZYMANSKI (LINEDATA)
- Lucien TANGHE (ASSURACTIS SARL)
- Eric TORDJEMAN (INRIA - INSTITUT DATAIA)
- Stéphanie TOUSSAINT (GRAND ENOV +)
- David TSANG-HIN-SUN (KEYRUS)
- Laura VELASCO AVISBAL (LABORATOIRES SERVIER)
- Eric VESSIER (ORACLE FRANCE)
- Richard VIDAL (ACCENTURE)
- Clément VIDON (SOCIÉTÉ CIVILE)
- Coline YVERGNIAUX (DEVOTEAM)

Initiative led by:

148, Bd. Haussmann - 75008 Paris - France
01 44 30 49 70 - contact@numeum.fr
@numeum

Supported by:

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre

3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

42

aivancity
SCHOOL FOR TECHNOLOGY, BUSINESS & SOCIETY
PARIS-CACHAN

ANITI
Université Fédérale

GRAND ENOV+

HUB FRANCE IA

IMPACT AI

INSTITUT DATAIA
Science des données, Intelligence & Société

Grand Est
ALSACE CHAMPAGNE-ARDENNE LORRAINE
L'Europe s'invente chez nous

Telecom valley
Animateur Azuréen Numérique

Conception graphique et réalisation : MAP ADVERTISING