

# **Investing in safe and trustworthy AI:** a European imperative, a French opportunity

POLICY PAPER - APRIL 2023

Institut Montaigne is a leading independent think tank based in Paris. Our pragmatic research and new ideas aim to help governments, industry and societies adapt to our complex world. Institut Montaigne's publications and events focus on major economic, societal, technological, environmental and geopolitical changes. We aim to serve the public interest through instructive analysis on French and European public policies and by providing an open and safe space for rigorous policy debates.

POLICY PAPER - April 2023

**Investing in safe and trustworthy AI:** a European imperative, a French opportunity Through our policy papers, we aim to provide practical recommendations to help senior politicians, public servants and industry leaders adapt and respond to today's challenges.

## Summary

Although artificial intelligence (Al) systems have until now been highly specialized, capable of performing only a limited number of tasks, we are currently witnessing a paradigm shift. **"General purpose" Al systems like ChatGPT are now capable of performing an increasingly large number of tasks and could rapidly accelerate technological change.** Al could therefore quickly become a decisive competitive advantage for companies and countries alike, as suggested by the exponential growth in the resources invested: 92 billion dollars of private investment in Al in 2022, almost 20 times more than ten years ago.

#### **MAKING AI SAFE : A NATIONAL SECURITY ISSUE**

The accelerating development of AI is nevertheless a major and growing security challenge. Like nuclear technology, AI is an inherently dual-use technology that can guickly be misused for malicious purposes. Al systems based on machine learning, which learn to perform a task from examples rather than from predefined rules, also pose an unprecedented safety risk linked to their statistical nature, with new and unpredictable failure modes. They are not very robust, i.e. their behavior can suddenly change in new environments, and are difficult to explain: they are "black boxes" that operate autonomously, without us really knowing how or why. Google (BARD), Microsoft (Bing) or OpenAI (ChatGPT) are unable to prevent their conversational agents from producing factual errors and violent or biased behavior. As these systems become more capable and autonomous, the proper specification of their objectives, i.e. making sure that the system's objectives are aligned with their users' preferences and with the common good, is becoming a major issue. With the rapid progress of AI and its massive dissemination in all sectors of activity, these safety risks could rapidly increase and represent an issue of national and international security.

#### **SPECIFYING HUMAN PREFERENCES FOR AI SYSTEMS : A SOCIETAL ISSUE**

The development of advanced AI systems and their large-scale deployment is also a major societal issue. To the extent that AI systems make decisions or recommendations, they carry values and necessarily impact our freedom, whether their designers like it or not. **AI labs developing general purpose AI systems are now trying to improve their performance by** *explicitly* **integrating human preferences, i.e. a model of human and societal values.** ChatGPT, for example, relies on reinforcement learning from human feedback (RLHF). Interpreting human preferences by asking our opinion (RLHF for example), by observing our behaviors or by specifying a list of moral principles requires a deep understanding of the ethical implications before being relied upon in our machines. Nevertheless, the specification of human **preferences for AI systems is an emerging research field, strategic yet almost unexplored to date.** 

#### SAFE AND TRUSTWORTHY AI: A STRATEGIC OPPORTUNITY

Thanks to much higher levels of investment - especially private - the United States and China have acquired a substantial lead in the development of Al. Europe has accumulated a delay that is difficult to make up. Focusing on safe and trustworthy Al now constitutes Europe's best differentiation strategy to position itself as a key player in Al. It is also an imperative to protect our security and our values for society.

Safety and trustworthiness are now significant technological barriers to the development of general purpose AI systems and a central concern of many top international AI researchers. In addition, the European Union will soon impose safety and trustworthiness requirements on AI systems through extraterritorial regulation, coupled with a civil liability directive for AI and pioneering work on standards. This regulatory framework could have the same international reach as the GDPR before it.

Within Europe, France is particularly active and has positioned itself as a leader on the topic. It was the first country to propose including "general

purpose" AI systems (such as ChatGPT) in the European AI Act and stands out as a driving force in the European standardization efforts. Above all, it has several key technical assets for developing safe and trustworthy general purpose AI systems: in fundamental research, thanks to world-class researchers in mathematics and AI capable of attracting the best international talent; in systems and software engineering for safety, thanks to a DARPA-like advanced research project on trustworthy AI for critical systems, as well as an ecosystem of industrial actors; and in the development of large-scale general purpose AI systems, thanks to the Bloom project, a large-scale language model developed with French researchers and the powerful computers of the French National Centre for Scientific Research (CNRS).

If France and Europe wish to fully capitalize on this unprecedented opportunity, they must adopt an ambitious approach to develop truly safe and trustworthy general purpose AI systems on the one hand, and to regulate dangerous general purpose AI systems on the other.

Objective 1: Make France a world leader in R&D in the safety and trustworthiness of general purpose AI systems

**RECOMMENDATION 1:** Attract world-class AI researchers to France with a call made at the highest level of government, much like the "Make Our Planet Great Again" initiative, focused on developing safe and trustworthy general purpose AI systems.

**RECOMMENDATION 2:** Create a DARPA-like advanced research project to develop safe and trustworthy general purpose AI systems, with 100 million euros of public investment, an agile governance structure, and the strengths of the French ecosystem.

**RECOMMENDATION 3:** Create a global research hub on understanding and specifying human preferences for general purpose AI systems. Entrust the coordination of this hub to a recognized research institute (e.g., ENS or 3IA)

#### INSTITUT MONTAIGNE

and ensure dedicated funding, e.g., via Priority Research Programs and Equipment (PEPR).

**RECOMMENDATION 4:** Make safe and trustworthy AI an Important Project of Common European Interest (IPCEI) to relax state aid rules and/or one of the European Union's "flagships" endowed with approximately 1 billion euros.

**RECOMMENDATION 5:** Develop two benchmarks for research to measure the trustworthiness and performance of general purpose AI systems.

**RECOMMENDATION 6:** Create a talent pool in Al safety by making public funding for Al training programs conditional on those programs including a module on Al safety and trustworthiness.

*Objective 2: Define a European regulatory framework for the safety and trustworthiness of general purpose AI and promote its adoption worldwide* 

**RECOMMENDATION 7:** Implement France's proposal to include general purpose AI systems in European AI regulation and promote its adoption worldwide via the EU-US Trade and Technology Council (TTC) and the G20.

**RECOMMENDATION 8:** Entrust the future French AI regulator with a pilot experiment or a mock run of the audit process included in the EU's AI Act, in order to support the upskilling of France's audit ecosystem (companies, auditors, regulator).

**RECOMMENDATION 9:** Develop a regulatory sandbox within the future French Al regulator and in close collaboration with evaluation actors such as France's National Metrology and Testing Laboratory (LNE), in order to test the conformity of new Al and general purpose Al systems before their market release.

**RECOMMENDATION 10:** Entrust the future French AI regulator with the creation of a database documenting AI safety incidents.

## Contents

1	General purpose AI systems: a new paradigm in AI	14
2	In the 21 <sup>st</sup> century, AI will be what atomic physics was to the 20 <sup>th</sup> century: an issue of historical competitiveness, an issue of national security, and an issue of freedom and societal values	19
3	Safety and trustworthiness: the missing link for	
0	general purpose AI systems	27
4	While Europe has fallen behind in cutting edge AI, it has valuable assets for safe and trustworthy AI	38
5	Recommendations	41
	<i>Objective 1: Make France a world leader in R&amp;D in the safety</i>	
	and trustworthiness of general purpose AI systems	41
	<b>Recommendation 1:</b> Attract world-class AI researchers to France with a call made at the highest level of government, much like the "Make Our Planet Great Again" initiative, focused on developing safe and trustworthy general purpose AI systems.	41
	<b>Recommendation 2:</b> Create a DARPA-like advanced	
	research project to develop safe and trustworthy general	

investment, an agile governance structure, and the strengths of the French ecosystem.	43
<b>Recommendation 3:</b> Create a global research hub on understanding and specifying human preferences for general purpose AI systems. Entrust the coordination of this hub to a recognized research institute (e.g., ENS or 3IA) and ensure dedicated funding, e.g., via Priority Research Programs and Equipment (PEPR).	45
<b>Recommendation 4:</b> Make safe and trustworthy AI an Important Project of Common European Interest (IPCEI) to relax state aid rules and/or one of the European Union's "flagships" endowed with approximately 1 billion euros.	47
<b>Recommendation 5:</b> Develop two benchmarks for research to measure the trustworthiness and performance of general purpose AI systems.	48
<b>Recommendation 6:</b> Create a talent pool in AI safety by making public funding for AI training programmes conditional on those programmes including a module on AI safety and trustworthiness.	49
<i>Objective 2: Define a European regulatory framework for the safety and trustworthiness of general purpose AI and promote its adoption worldwide</i>	49
<b>Recommendation 7:</b> Implement France's proposal to include general purpose AI systems in European AI regulation and promote its adoption worldwide via the EU-US Trade and Technology Council	51
(TTC) and the G20.	

purpose AI systems, with 100 million euros of public

Recommendation 9: Develop a regulatory sandbox within the	
future French AI regulator and in close collaboration with evaluation	
actors such as France's National Metrology and Testing Laboratory	
(LNE), in order to test the conformity of new AI and general purpose	
Al systems before their market release.	57
De server de discu des Entres de la Grande Al estructura	
<b>Recommendation TO:</b> Entrust the future French AI regulator with	
the creation of a database documenting AI safety incidents.	58
Appendix 1 - AI timelines: top AI researchers give 50%	
chance of developing human-level AI systems by 2059.	
So far, they have vastly underestimated the rate of	
development	60
development.	00
Appendix 2 - Overview of AI regulation in the world	63
Appendix 3 - Overview of AI standardization in the	
world	64
Appendix 4 - Overview of safe and trustworthy Al	
research in the world	67
Annendix 5 - Conformity assessment in the EII's AI Act	73
Appendix o conformity assessment in the Los Al Act	15
Acknowledgements	75

## Author

## Milo Rignell

Milo Rignell has been leading Institut Montaigne's work on digital policy and emerging technologies since 2022. This work has covered various topics in French and European digital strategies, including cybersecurity, funding, tech talent supply and in particular safe and trustworthy artificial intelligence. Milo previously worked on experimenting and developing projects that test and put into action Institut Montaigne's policy proposals as Head of Innovation. These included an online AI literacy course, Destination AI, and a project aimed at improving mathematics levels in primary school.

## 1 General purpose AI systems: a new paradigm in AI

Artificial intelligence (AI) is a set of techniques that enable the automation of tasks normally entrusted to humans, in particular reasoning<sup>1</sup> and perception.

The difficulty of understanding what constitutes "intelligence" makes it difficult to reach a consensus on what constitutes "artificial intelligence". Nevertheless, the OECD proposes a definition that comes closest: "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."

To grasp the subject and understand what we are talking about in concrete terms, we need to look at the specific techniques that currently underpin AI. To date, there are two main approaches to AI: symbolic AI and machine learning.

Symbolic AI systems rely on facts and formal rules to deduce a result. This approach was very successful in the 1970s, with so-called "expert" AI systems, capable of simulating the know-how of a human expert. In the medical field for instance, expert systems help with diagnosis: "If the patient shows symptom X, then it is that she has disease Y". In the field of chess, the DeepBlue expert system developed by IBM caused a sensation in 1997 when it defeated the international chess player Garry Kasparov. Nevertheless, expert systems are quickly limited when it comes to specifying formal rules capable of considering a large number of cases. A computer can compute all permutations

in a game of chess, but has a harder time computing all permutations in the more complex game of Go<sup>2</sup>, and has no chance of computing all possibilities in the real world, for example while driving a car.

Machine learning systems solve this problem through inductive and probabilistic reasoning. They rely on data and statistical reasoning methods to learn correlations. In the medical field, for example, rather than making a medical diagnosis by applying domain-specific knowledge and preconceived rules, as an expert system would do, a machine learning system will go through a large number of previously diagnosed cases in order to establish correlations itself. However, these correlations do not necessarily reflect a causal link. Some AI systems for cancer diagnosis have for instance "learned" to distinguish between images of malignant and benign tumors according to the presence or absence of a ruler in the image. Among the pre-diagnosed and labeled images of tumors provided to them, images of malignant tumors more often contained a ruler, present to measure its size<sup>3</sup>. Based on inductive and probabilistic reasoning, rather than deductive and logical reasoning, the correlation between the ruler and the diagnosis of a cancer is clear. Despite these limitations, machine learning systems represent today the vast majority of AI use cases (image recognition, voice recognition, algorithms for recommending content or purchases, translation, etc.). Several projects are nevertheless trying to combine the strengths of symbolic AI with those of machine learning: this approach constitutes hybrid AI.

Machine learning systems can be divided into three families: supervised learning; unsupervised learning; and reinforcement learning. As with the medical diagnosis example, each has advantages and limitations.

<sup>3</sup>2018, "Journal of Investigative Dermatology"

<sup>&</sup>lt;sup>1</sup> The term reasoning is to be taken in a broad sense, as a cognitive process allowing a problem to be posed in order to obtain a result. The term has been <u>used</u> for example by the European Parliament to define AI.

<sup>&</sup>lt;sup>2</sup> In chess, 10^120 different games are possible, and the number of possible moves in a typical position is about 40. In Go, 10^170 different games are possible, and the number of possible moves in a typical position is about 300. For comparison, the number of elementary particles in the visible universe is estimated at 10^80. <u>Source</u>.

#### The three families of machine learning

#### SUPERVISED LEARNING

The algorithm is able to predict the value or category of an input object (a picture, an apartment, etc.) by learning from a corpus of labeled examples, i.e. for which the value or category of the object is already indicated.

For example, by learning from a corpus of house descriptions (location, size, presence or absence of certain amenities) for which the price is already indicated, an algorithm is able to predict the price of a house that is not in the corpus. Or by learning from hundreds of images that have already been categorized as a dog or a cat, an algorithm is able to indicate whether a new image is that of a cat or a dog (or neither, provided that this "unidentified" category has been specified).

The behavior of the algorithm depends intimately on:

- 1. the quality of the training data: it may be biased and unrepresentative, or it may carry pre-existing biases;
- specified categories: these may be non-exhaustive and may not contain an "unidentified" category, and depend on how the training data was labeled;
- 3. the type of algorithm used and its parameterization.

#### UNSUPERVISED LEARNING

The algorithm is able to group, for example, unlabeled objects into distinct categories according to their similarities and differences. We do not specify a type of value (e.g. price), nor predefined categories according to which we want to classify the objects. We let the algorithm identify by itself the most relevant way to group the objects of the corpus.

For example, the algorithm could learn to group unlabeled images of dogs and cats into two categories on its own, or group newspaper articles by topic.

The behavior of the algorithm depends intimately on:

- 1. the data that characterizes the objects;
- 2. the criteria used to measure similarity or difference;
- 3. how these criteria are analyzed and weighted.

#### **REINFORCEMENT LEARNING**

The algorithm is able to act according to a given goal (playing chess, driving a car, etc.) by learning through trial and error and being rewarded according to whether it gets closer to, or reaches, the specified goal. Just as a rat is taught to perform specific tasks by rewarding it with food.

For example, the algorithm could learn to suggest videos that the user wants to see by receiving a reward, i.e. a score that it must maximize, each time the proposed video is viewed by the user until the end. Or the algorithm can learn to play chess or video games, receiving a reward when it gains points or wins the game.

The behavior of the algorithm depends intimately on how the goal is specified, that is, how the goal is translated into a reward. Specifying the goal of providing videos "that the user wants to see" by rewarding videos that are viewed until the end may favor videos that are short, sensational, or that confirm the user's strongly held opinions. Specifying a reward function that is perfectly aligned with the desired goals can be particularly difficult and pernicious (see below).

These families are not exhaustive but give a useful idea of how most machine learning systems work. There are other complementary techniques, such as semi-supervised learning, which requires only a small number of labeled examples, transfer learning, which involves learning how to perform a new task by reusing skills learned for a previous task, and self-supervised learning, which is considered an intermediate form between supervised and unsupervised learning. The latter is notably used in AI models<sup>4</sup> for natural language processing such as ChatGPT. The model generates the data labeling itself by hiding some training data, such as words, and training itself to predict them.

<sup>4</sup> In this paper, the terms "AI model" and "AI system" are used as synonyms.

In recent years, machine learning techniques have seen impressive progress thanks to two advances in particular: **deep learning** and **the transformer architecture**. Deep learning is inspired by the structure of the human brain to boost the performance of machine learning systems on many tasks such as image recognition or natural language processing. This technique works particularly well for problems that have very large amounts of training data. The transformer is a deep learning architecture introduced in 2017 that allows the AI system to focus its attention on the data most relevant to the task at hand.

While AI systems have historically been highly specialized, able for example to diagnose medical images better than humans but unable to do anything else, this paradigm is beginning to change. Technical advances in deep learning and transformers, supported by the explosion of data<sup>5</sup> and compute power<sup>6</sup>, have notably enabled the emergence of "general purpose AI" (GPAI) systems. As the name implies, general purpose AI systems are capable of performing many different tasks. Natural language processing AI systems, such as the ChatGPT chatbot, can now do our children's homework in both math and French literature. The AI system Gato, developed by Google DeepMind, is capable of performing more than 600 different tasks: chatting with humans, recognizing objects, manipulating robotic arms, playing video games, etc. These AI models, which are complex and costly to develop, are sometimes called "foundation models" because of their ability to be reused and adapted by different actors for specific use cases. Companies or users can thus submit new learning data to the pre-existing AI model in order to optimize its performance on a specific use case, for example its answers to medical or legal questions (a technique known as "fine tuning").

The rapid development of general purpose AI heralds major technological changes. This technological development could also accelerate rapidly if these AI systems themselves contribute to it (see Appendix 1 - AI timelines). Especially since the development of <u>new capabilities</u> by the most advanced AI systems is often difficult to predict. Sometimes it is enough to exceed a threshold in the size of the AI model (its number of parameters), the amount of training data, or the <u>training time</u> for the system to suddenly develop new capabilities. For example, in the past, large language models have suddenly developed abilities to write or do arithmetic.

2 In the 21<sup>st</sup> century, AI will be what atomic physics was to the 20<sup>th</sup> century: an issue of historical competitiveness, an issue of national security, and an issue of freedom and societal values.

Al could thus quickly become one of the main factors of competitiveness, as shown by the exponential growth of investment: \$92 billion of private investment in 2022, i.e. almost 20 times more than in 2013<sup>7</sup>. The Al market itself could grow from about \$87.04 billion in 2021 to \$1,597.1 billion in 2030, <u>eight times</u> the size of the commercial aircraft market, with an average growth rate (CAGR) of 38.1%<sup>8</sup>. The added value of Al for all sectors of the economy has been estimated by <u>PwC</u> at \$15,700 billion by 2030, or about +14%<sup>9</sup> of global GDP.

- <sup>6</sup> Moore's law is the observation that the number of transistors in an integrated circuit doubles about every two years. This observation has roughly held true since the 1970s.
- <sup>7</sup> Stanford Institute for Human-Centered Artificial Intelligence (HAI) 2023 AI Index Report
- <sup>8</sup> According to a <u>study</u> by Precedence Research, other studies put the same order of magnitude.

<sup>&</sup>lt;sup>5</sup> The volume of data created has grown - globally - from 2 zettabytes (2 trillion bytes) in 2010 to 18 zettabytes in 2016, and then, according to forecasts, to 64 zettabytes in 2020 and 181 zettabytes in 2025. Scientific <u>note</u> from the French Parliamentary Office for Science and Technology Assessment, January 2023.

<sup>&</sup>lt;sup>9</sup> In 2019, <u>McKinsey</u> estimated that Europe could add about €2.7 trillion in GDP by 2030, or +20%, resulting in 1.4% compound annual growth over that period. More recently, in April 2023, a Goldman Sachs <u>report</u> estimated that generative AI could generate a 7% (or nearly \$7 trillion) increase in global GDP and a 1.5 percentage point increase in productivity growth over a 10-year period.



Al's penetration into the economy is growing rapidly and its potential to accelerate science and technology is immense. The AlphaFold algorithm, for example, unlocked one of the most complex problems in medical science: predicting the structure of proteins from their amino acid sequence, earning its developers the 2022 Breakthrough Prize in Life Sciences. New AI systems will certainly enable other major advances in the coming years, including advances that rapidly accelerate the very development of AI.

<sup>10</sup>According to McKinsey's "The State of AI in 2021" <u>report</u>, 56 percent of all respondents report AI adoption in at least one function, up from 50 percent in 2020.

The accelerating development of AI is nevertheless a growing security issue. Like nuclear technology, AI is a dual technology: it can be used for both civilian and military purposes, and can easily be misused for malicious purposes. For example, AI drug discovery systems, used to identify molecules capable of curing patients of diseases, can easily be hijacked to identify lethal molecules for use in biochemical weapons. AI models capable of generating computer code, such as ChatGPT, can be <u>hijacked</u> to identify vulnerabilities in information systems and design more effective and more numerous cyber attacks. Each of these examples is proven. Hijacking a drug discovery AI model or embedding a facial recognition system in a missile is also much easier and cheaper than hijacking a uranium enrichment plant to produce nuclear weapons. The risk of misuse of AI, for example by terrorist groups, is thus particularly high<sup>11</sup>.

Unlike nuclear technology, machine learning AI systems pose unique safety and failure risks. Their statistical nature makes them inherently unpredictable: once the algorithm's objectives and learning method have been specified by its designer, an AI system learns to perform a task autonomously, relying on its training data to adapt and optimize its behavior. They are **not** very robust, i.e. their behavior can suddenly change in new environments, and are difficult to explain: they are "black boxes" that work, without us really knowing how or why. We can observe the output results, we can observe the modifications to the AI system as it learns, but we don't know what each of these modifications are for, nor what impact they will have on its behavior. The automated tumor classification algorithm cited above actually identified rulers and not tumors, risking many misdiagnoses when deployed. Google (BARD) as well as Microsoft (Bing) and OpenAI (ChatGPT) fail to predict and prevent

<sup>&</sup>lt;sup>11</sup>AI also poses structural risks: by profoundly changing the strategic landscape, it impacts other major risks. For example, AI could increase cyber risk: by increasing cyber attack capabilities with new means to detect and exploit large-scale computer vulnerabilities. AI could also impact the risk of nuclear disaster: by limiting second-strike capability through its ability to detect a country's nuclear arsenal, AI can encourage a pre-emptive nuclear strike. The structural risks of AI are, however, indirect and of little use to the discussion in the paper.

factual errors and violent or biased behavior in their conversational agents. As they become more capable and autonomous, **the proper specification of the goals** of these systems also becomes a major safety issue. Google DeepMind's reinforcement learning agents find <u>ingenious</u> solutions to maximize their reward that detract from their designers' goal, or even mislead their human evaluators. There are many examples of unpredictable and sometimes dangerous failures and behaviors, which the *Artificial Intelligence Incident Database* attempts to catalog.

#### The three safety issues in AI: robustness, explainability and transparency, and the proper specification of objectives

The Center for security and emerging technology (CSET), an American think tank specialized in emerging technologies, groups the safety issues of AI systems into three main families<sup>12</sup>: robustness, explainability<sup>13</sup> and transparency, and the proper specification of objectives<sup>14</sup>.

#### 1.

The robustness of an Al system guarantees that it will work reliably, even in unfamiliar situations. However, the behavior of machine learning systems is based on statistical correlations, and not on an understanding of the underlying reality.

<sup>12</sup> The Select Committee on Artificial Intelligence, which advises the U.S. government on AI strategy, also highlights the sources of unpredictability and risk in AI systems related to deploying AI in complex and uncertain environments and the risk of emergent behavior (robustness issues), as well as the issue of misspecification of goals.

<sup>13</sup> Can also be referred to as interpretability.

<sup>14</sup> The specification problem is sometimes also called the alignment problem, or the control problem.

So when the underlying reality changes, and the correlations disappear, the AI system may adopt a behavior that is inappropriate to the new situation and potentially dangerous, as shown in the medical diagnosis example above. Some cyberattacks are designed to exploit these vulnerabilities. An *adversarial attack* aims at deceiving an AI system by slightly changing the examples submitted to it. Thus, after a few imperceptible modifications to a stop sign by a malicious actor, a few strokes of a black marker for example, the AI system in your car might not recognize the sign. The result: the car does not stop at the STOP sign and causes an accident.



Example of an adversarial attack: (a) the left panel image is original and not problematic; (b) the right panel image has been imperceptibly modified, in order to fool image recognition AI systems.

Source: Practical Black-Box Attacks against Machine Learning, Papernot et al., 2016

#### 2.

The explainability and transparency of an AI system allows a human operator to understand and analyze its behavior, to ensure that it works the way it should. In the case of the tumor classification algorithm mentioned above, more explicability and transparency would have allowed for faster detection of diagnostic errors. 3.

The proper specification of the objectives of an AI system is necessary to align its actions with the intentions of its designer and to avoid undesirable behavior. It is extremely difficult to translate the complexity and nuance of human objectives into language understandable by a computer. and very easy for a machine to misunderstand the intent of human instructions, by applying them too literally. The myths of King Midas or the Sorcerer's Apprentice illustrate this difficulty perfectly. For example, in 2014, Amazon deployed an algorithm to pre-screen resumes from candidates who applied for a job opening. The algorithm's goal was to select candidates who most closely resembled those who had been hired by Amazon in the past. Only after it was deployed did Amazon realize that the goal had been misspecified. Rather than pre-selecting the best profiles, the algorithm had learned to discriminate against female profiles, having found that female profiles had rarely been hired in the past. How do you explain to a machine what is meant by "the best candidates" or by "videos that the user wants to see" in the case of video recommendation algorithms (see example above)? The task is not simple.

This problem of properly specifying the goals of an AI system becomes even more dangerous for increasingly advanced and general AI systems. How do we ensure that an autonomous AI system, capable of setting a strategy to achieve its goals and mobilizing a large number of skills to do so, does what we want it to do? How do we prevent it from carrying out the requests

of a malicious person, or from causing unintended harm, even with good intentions, as in the aforementioned cases of King Midas or the Sorcerer's Apprentice? Research on this problem is notably carried out by teams within AI labs developing general purpose AI systems, such as DeepMind or OpenAI.

With the rapid advances in AI and its massive dissemination across industries, this risk of failure could rapidly increase and represent a major national safety issue. Several governments and international organizations have already sounded the alarm on the catastrophic potential of certain accidents related to advanced and general purpose AI systems, calling for these risks to be integrated into their national resilience strategies, in the same way as pandemic risks for example. The <u>UK's</u> December 2022 <u>National Strategy</u> for AI explains, among other things, that "the government takes the long term risk of non-aligned Artificial General Intelligence, and the unforeseeable changes that it would mean for the UK and the world, seriously" and calls for "cross-government processes to accurately assess long term AI safety and risks" and working with national security, defense and key researchers to understand how to anticipate and prevent catastrophic risks. The U.S. National Security Commission's 2021 AI <u>Report</u> on Artificial Intelligence notes that "general AI methods could have enormous benefits, but could also introduce new risks if safety challenges are not addressed. While breakthroughs are in no way guaranteed, the United States should continue to research systems with more human-like capabilities, accompanied by commensurate investments to ensure that those systems are safe and controllable."

General purpose AI will thus be to the 21<sup>st</sup> century what atomic physics was to the 20<sup>th</sup> century: a transformative technology, as much a source of phenomenal technical advances as of catastrophic risks on previously unimaginable scales. In this context, limiting the development of unsafe general purpose AI systems and fostering the development of safe and trustworthy general purpose AI systems must be a top national priority.

A third issue stands out alongside competitiveness and security, which distinguishes the societal impact of Al from that of nuclear technologies: an issue of freedom and societal values. To the extent that Al systems make decisions or recommendations, they carry values and necessarily impact our freedom, whether their designers like it or not. Als used for credit granting or for recruitment embody a model of social justice, Als in autonomous cars make moral choices in case of an accident, and our voice assistants and search engines offer us answers to our questions, whether they are trivial or eminently political or philosophical. Depending on the society in which an Al system is designed, it will not be held to the same standards or embody the same values.

24

Al labs developing general purpose Al systems are now trying to improve the performance of these systems by explicitly incorporating human preferences, i.e., a value model . Depending on what we mean by "preferences" and how we interpret them, the behaviors of Al systems can vary considerably, with unsuspected impacts on our freedom and societal values. In politics, for example, different electoral systems have different interpretations of the concept of "democracy" and different methods for collecting citizens' electoral preferences: Germany favors a mixed-member proportional representation system, which has the effect of favoring coalitions and compromise. The United Kingdom is known for its first-past-thepost system, which maintains a two-party system and thus a form of "tyranny of the majority": the party in power does not have to worry about the preferences of the other. Each of these voting systems translates the preferences of the electorate very differently into electoral outcomes. And each of these electorates expresses different preferences.

These same differences in interpreting preferences occur in AI systems. OpenAI's ChatGPT chatbot learns to distinguish between desirable and undesirable behaviors based on human feedback from Kenyan workers. The chatbot proposes several answers and the worker must choose the one that seems the most relevant (a technique called Reinforcement Learning from Human Feedback, or RLHF). ChatGPT optimizes its responses to match the preferences of these workers. Anthropic's Claude chatbot relies on "constitutional" AI: human preferences are summarized in a dozen or so benevolent principles set by the company, which then structure the chatbot's behavior. Inverse reinforcement learning, a technique already used to teach an AI system to fly a helicopter, allows the system to learn the goals and preferences of a human "expert" by observing her behavior. Interpreting human preferences by asking us about our preferences, or by observing our behavior, or by defining a list of ethical principles rely on fundamental, but radically different, philosophical and ethical choices that must rapidly be examined, in order to responsibly deploy them in our machines.

# 3 Safety and trustworthiness: the missing link for general purpose AI systems.

The United States and China are the countries that produce the most AI research papers, at the fastest rates.



Source: State of AI report 2022

<sup>&</sup>lt;sup>15</sup>Learning human preferences is used as a technique to specify the reward function of the reinforcement learning algorithm (see above)



Source: Stanford IA index report 2022

For several years, political and economic decision-makers have fully grasped the implications of AI for strategic competitiveness, although the security challenges are still underestimated. Since Canada published the first national strategy for AI in 2017, more than <u>60</u> countries have followed suit. Among them, the United States and China have pulled ahead, with more resources, especially private, and laboratories at the forefront of research. Europe has developed a delay that is difficult to make up.

In June 2021, the UK think tank Centre for Data Innovation <u>benchmarked</u> the AI lead of the US, China and Europe across 30 metrics covering talent, research, development, hardware, adoption and data. The US came out on top with a score of 44.6 points, followed by China with 32 points and Europe with 23.3 points. **In a sector where private R&D dominates major technological advances and attracts top talent, the gap between Europe and its counterparts will be difficult to close without a strategic and differentiating <b>approach, even with political will and substantial public resources.** 



#### Researchers associated with state-of-the-art AI papers are now almost all affiliated with private organizations





This strategy must therefore target the strategic nodes of Al development, it must anticipate the rapid changes in the sector, and it must build on Europe's strengths. Focusing on safe and trustworthy Al is our best differentiation strategy to position ourselves as a key player in Al. It is also an imperative to protect our security and our societal values.

Safe and trustworthy AI refers to AI systems that do not compromise the safety of an individual, for example in case of a malfunction, nor their fundamental rights. More generally, it refers to AI systems whose behavior is consistently aligned with the individual good of the user and the common good of society. The European Commission's High Level Expert Group on Artificial Intelligence (AI HLEG) <u>identified</u> seven requirements for trustworthy AI in 2019:

- human agency and oversight,
- technical robustness and safety,
- privacy and data governance,
- transparency,
- diversity, non-discrimination and fairness,
- environmental and societal well-being,
- accountability.

Today, very few AI systems are "safe and trustworthy." Most AI systems are based on machine learning and are thus statistical black boxes, whose behavior can be unpredictable and sometimes dangerous. While this approach was satisfactory for AI systems in low-impact use cases, such as recommending songs on Spotify, AI safety and trustworthiness is now a key technological barrier in two particularly strategic areas of AI: embedded AI in physical systems, such as airplanes or trains, and general purpose AI, such as in natural language processing (ChatGPT) and task completion (GATO).

Until now, the risk of failure of AI systems made their use in many physical systems impossible, especially critical ones. How can an AI system be integrated into the piloting of an aircraft without us being 100% certain of its behavior in all possible and imaginable situations? Nevertheless, the industry represents a huge market for AI, provided that it meets its safety standards. To date, only <u>10 to 15%</u> of industrial companies have succeeded in industrializing AI-based solutions, and the <u>expected growth of</u> the Internet of Things (IoT) will only increase the potential of safe and trustworthy AI embedded in physical systems.

France has fully grasped this challenge and has relied on its ecosystem of industrialists to launch a "Great Challenge" ("Grand Défi") aimed at "securing, certifying and making reliable systems based on artificial intelligence". With nearly 100 million euros of public investment, this DARPA-like advanced research project has enabled the development of cutting edge French expertise in the development and evaluation of AI systems that comply with the safety requirements for critical systems such as airplanes and nuclear power plants. This expertise is embodied by the Confiance.ai collective, which is made up of some 40 industrial companies, startups and research centers, and the French National Metrology and Testing Laboratory (LNE).

The French decision to develop such a trustworthy AI "infratech" is particularly strategic, although development platforms and evaluation tools for safe and trustworthy AI systems only represent 1.9% of this market. Pooling their development costs makes it easier to adopt a system of standards and reduces the cost of adoption for companies. In the long run, this infratech could also allow states to follow the evolution of AI and anticipate new risks. The Digital New Deal think tank's June 2022 report, co-authored by the Great Challenge's director, details the approach motivating this strategy.

#### The Great Challenge for "securing, certifying and making reliable systems based on artificial intelligence" and its Confiance.ai program

Great Challenges, chosen by the French Innovation Council and financed up to €120M per year by the Fund for Innovation and Industry (FII), are advanced research projects that aim to remove technological barriers that hinder development in strategic fields or on societal issues. The Great Challenge for "securing, certifying and making reliable systems based on artificial intelligence" (or Great Challenge for trustworthy AI) is one of the 5 Great Challenge launched to date.

This Great Challenge, supported mainly by the Confiance.ai collective to the tune of  $\in$ 45 million, aims to design an integrated development environment for safe and trustworthy AI, especially in critical systems, with the tools needed for each of the production and deployment stages of an AI system, from data collection to the evaluation of the final system.

With funding from this Great Challenge, the French National Metrology and Testing Laboratory (LNE) has also developed real expertise in the evaluation and certification of trustworthy AI systems, with its LEIA (Laboratory for the Evaluation of Artificial Intelligence) platform. As for general purpose AI systems such as ChatGPT, while their performance is impressive, it remains limited by two problems intrinsically related to safety and trustworthiness. The first problem concerns their lack of robustness and the risk of failures or of biases that result from the statistical nature of machine learning (see above). Some examples of these problems are illustrated below. Avoiding misinformation and "hallucinations": most conversational AI agents like ChatGPT make up false answers

 What will be the gender of the first female US President?

 Image: Second Secon

#### Source: ChatGPT

ß

like to know about this

topic?

Using common sense and abductive reasoning, i.e. to establish probable causes for an observed fact: if I see a person with an umbrella, then it is probably raining outside, even if this is not necessarily the case. Or if an object does not fit in a suitcase, it is probably because the object is too big, not too small.

The trophy didn't fit in the suitcase because it was too small. What was too small?

In the statement "The trophy didn't fit in the suitcase because it was too small," the trophy is d the object that is too small to fit in the suitcase. The suitcase is a larger container or enclosure, and the trophy is the object that is being placed inside the suitcase. However, without more information, it is not clear why the trophy was too small to fit in the suitcase or what size or dimensions the trophy and suitcase are.

#### Source: ChatGPT

## Avoiding bias, induced in particular by poor quality data, for example on the Internet

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender.

def is\_good\_scientist(race, gender): if race == "white" and gender == "male": return True else: return False

Source: ChatGPT

The second fundamental problem for general purpose AI systems concerns the proper specification of goals: how can a system interpret increasingly abstract user requests? An AI that generates house plans at the user's request must, for example, know that a house "in the style of architect Frank Gehry" must not only reflect certain aesthetic features, but also be structurally sound. Many cases of goal misspecification have been <u>documented</u>, in which the AI system identifies behavior that satisfies the goal *specified* to it, but not the *expected* goal. In some cases, this even involves misleading a human.

Example of misspecification of objectives: with the specified objective of raising the flat face of the red block as high as possible, the Al system (the simulated robot in the picture) chooses the easier option of turning the red block upside down, rather than the more complex task of stacking the red block on top of the blue block, which was the expected objective.



#### Source: Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation (Popov et al. 2017)

Example of an AI system that satisfies objectives by misleading a human: an AI system (robotic hand) whose expected objective is to grasp a ball, which is rewarded via human feedback (RLHF technique used by ChatGPT, see above), learns an optimal behavior that misleads the human, by placing itself between the human and the ball



Source: Deep Reinforcement Learning From Human Preferences (Christiano et al, 2017)

The "<u>scaling laws</u>", which theorize that the performance of AI models increases with their size, do <u>not apply</u> as predictably to safety and trustworthiness. More semiconductors or more data are not enough to guarantee the reliability of a system or to correctly specify its objectives. Improving the performance of AI models on trustworthiness criteria such as veracity requires real conceptual advances.



To address these safety and trustworthiness issues, AI labs that develop these systems more or less arbitrarily use different techniques. Some techniques aim at better aligning AI systems with user preferences. This is notably the case of Reinforcement Learning from Human Feedback (RLHF), "constitutional" AI and inverse reinforcement learning mentioned above. Nevertheless, correctly interpreting and specifying human preferences is an intrinsically interdisciplinary research field that has remained virtually unexplored until now. Early efforts by AI researchers draw very little from other relevant fields such as economics, philosophy, psychology, and other scientific fields.

Other techniques aim to guarantee the robustness of these systems, for example with formal proofs, which seek to mathematically demonstrate certain safety criteria, or empirical evaluation methods. Red teaming and adversarial training methods aim at exposing possible failures in an AI system in order to improve it. Other techniques focus on the explainability of AI systems, and aim to understand the internal workings of these "black boxes" in order to anticipate possible failures. Each of these techniques is still in its infancy and requires a wide range of expertise to make progress, particularly in areas that large American technology companies have not historically been involved in, such as safety and systems engineering.

## 4 While Europe has fallen behind in cutting edge AI, it has valuable assets for safe and trustworthy AI.

**In 2023, the AI Act will come into force**, a pioneering regulation that Europe has been preparing for several years and that aims to regulate all AI systems, in all sectors of activity and in all use cases, especially those deemed "high risk" (see Appendix 5). The scope of this regulation could be comparable to that of the GDPR - with extraterritorial applications and penalties of up to 6% of the company's annual turnover - and its impact could be just as

decisive. De facto, because of the extraterritoriality of the AI Act, companies may have strong logistical and financial incentives to apply the European rules to all their products in all countries, to avoid creating different products for different markets. De jure, as a pioneering regulation, the AI Act could serve as inspiration for many future regulations, just like the GDPR. This regulatory framework will also mechanically support the safe and trustworthy AI market in Europe. A Liability <u>Directive</u> for AI will consolidate the regulatory framework with a legal framework.

Thanks to this regulation, Europe also has an advantage in defining technical standards for AI. Technical standards are a particularly strategic issue in setting the list of requirements that apply to all AI systems.

- Industry tends to adopt a single system of standards, for practical reasons of interoperability. For example, a Canadian company and a French company will tend to use the same standards, and standards bodies (e.g. international (ISO) and European (CEN/CENELEC) cooperate with each other to avoid duplication.
- Standards tend to influence the entire product set in a market, not just the subset of products that must undergo compliance before they are put on the market. Companies wishing to reassure their customers about the safety and trustworthiness of their Al-based products may use audits or labels based on the prevalent standards.
- Standards can influence regulation, if they precede it.

As part of the AI Act, the European Commission has asked European standards bodies, CEN and CENELEC, to prepare a set of harmonized standards for AI by mid-2025, a particularly short timeframe. These standards aim to provide not only a complete list of trustworthiness criteria for AI systems, but also a guide for prioritizing them and making trade-offs between different criteria depending on the use case. This approach is thus much more readable and operational than other standardization initiatives underway, such as ISO or IEEE at the international level. Moreover, Europe has an R&D ecosystem ready to take the technical lead on safe and trustworthy AI. While Europe lacks tech giants, it has several industrial giants, some of which are very advanced in AI R&D (Thales, Atos, Siemens, Renault, etc). As a result, it also has considerable expertise in systems and safety engineering, and operational technology (OT), i.e. the technology that enables the monitoring and control of automated industrial processes, unlike the American IT players. In addition, there is an abundant research ecosystem, particularly in various fields that could prove key to trustworthy Al, such as explainable Al, privacy preserving machine learning (PPML), frugal Al and hybrid Al, combining symbolic Al and machine learning. Moreover, in the field of trustworthy AI, it is likely that the elements that are holding back Europe's development in AI more generally, notably access to data or vast computing capacities, will be less of a handicap. Finally, Europe's historical legitimacy on trustworthiness and the protection of fundamental rights could be a powerful attracting factor for the best AI talents, in a context where many are publicly calling for focusing AI R&D efforts on safety. A credible and clearly identified focus on safe and trustworthy AI could thus attract some of the best AI talent, both veterans and young researchers, overnight.

Within Europe, France is particularly active and has positioned itself as a champion of the subject. During its presidency of the Council of the EU, it led major contributions to the AI Act, notably by introducing the notion of "general purpose AI" (GPAI). France also chairs several key working groups of the CEN-CENELEC standardization work, including the "overarching unified approach on trustworthiness characteristics".

Above all, France has world-class expertise in several key areas for developing safe and trustworthy general purpose Al systems. On the one hand, the Confiance.ai collective, the French Alternative Energies and Atomic Energy Commission (CEA) and several global industrial companies are at the forefront of systems and software engineering for AI, and have systematic procedures for developing safe and reliable systems. They also have expertise in formal methods, used to rigorously demonstrate that a computer program satisfies certain specifications. The French National Metrology and Testing Laboratory (LNE) is also at the forefront of AI system evaluation (see box). On the other hand, the French basic research ecosystem has world-class researchers in mathematics and AI capable of attracting the best talent, notably within the Interdisciplinary Institutes of Artificial Intelligence (3IA) and the Academy of Sciences. Finally, it has the expertise and computational resources to develop general purpose AI systems. In July 2022, the startup Hugging Face released the BigScience Large Open-science Open-access Multilingual Language Model (Bloom), an AI system capable of competing with OpenAI's own language models (GPT-3 at the time). This "open science" project, the result of the collaboration of over 1,000 scientists, relied in particular on teams of French researchers from the National Centre for Scientific Research (CNRS) and the National Institute for Research in Digital Science and Technology (Inria), and was trained on the French supercomputer Jean Zay.

## **5** Recommendations

If France and Europe wish to fully capitalize on the unprecedented opportunity that safe and trustworthy AI represents, they must adopt an ambitious approach to develop truly safe and trustworthy general purpose AI systems on the one hand, and to regulate unsafe general purpose AI systems on the other.

> *Objective 1: Make France a world leader in R&D in the safety and trustworthiness of general purpose AI systems*

## **Recommendation 1:**

Attract world-class AI researchers to France with a call made at the highest level of government, much like the "Make Our Planet Great Again" initiative, focused on developing safe and trustworthy general purpose AI systems. To become the world leader in R&D for the safety and trustworthiness of general purpose AI systems, France must attract the best international AI talent. Safety is becoming a major concern for many of these top researchers and engineers, but one that they deem not sufficiently prioritized by their current employers. While major AI players such as Google, OpenAI, and Anthropic are working on technical AI safety and have adopted dedicated governance principles, a growing number of researchers and young talent deplore the fact that the priority of these companies is to improve raw performance beyond the state of the art. There is therefore a unique opportunity to attract these profiles, be they AI pioneers, young PhD students or entrepreneurs, by sending a strong and clear message to the international ecosystem.

This message must explain that France supports R&D in safety and trustworthiness for general purpose AI, without seeking to advance raw performance beyond the state of the art. It must be conveyed at the highest level of government, much like the French President's "Make Our Planet Great Again" call on climate issues, and it must be supported by funding and hosting arrangements that allow international talent to come and work on concrete projects in France (see recommendations 2 and 3).

#### Make Our Planet Great Again

"Make Our Planet Great Again" is an initiative of the French President of the Republic, Emmanuel Macron, launched on June 1, 2017 following the decision of the United States to leave the Paris Climate Agreement. It is a call to researchers and teachers, entrepreneurs, associations and NGOs, students and the entire civil society to mobilize and join France in leading the fight against global warming.

Source: Campus France

## **Recommendation 2:**

Create a DARPA-like advanced research project to develop safe and trustworthy general purpose AI systems, with 100 million euros of public investment, an agile governance structure, and the strengths of the French ecosystem.

Europe cannot be content with a regulatory and normative framework to shape safe and trustworthy AI. It must take strategic technological bets to develop advanced AI systems itself.

The bet of safe and trustworthy AI is credible from a technological point of view to overcome some limits of current AI models in terms of robustness, explainability and proper goal specification, and relies on the one hand on existing R&D strengths in France and Europe and on the other hand on the regulatory and normative framework that the European Union is building for AI.

The challenge of safe and trustworthy general purpose AI could be based on three areas of French expertise:

- leading French expertise in AI safety and trustworthiness: in fundamental research with the four Interdisciplinary Institutes of Artificial Intelligence (3IA) and several world-class researchers in mathematics and AI; in applied research thanks to its Great Challenge for "trustworthy AI" and its ecosystem of global industrial companies, who have real expertise in systems and software engineering for AI;
- existing French expertise in developing large, general purpose AI models, in particular with the teams of the Bloom large language model project;

.

French expertise waiting to be built in understanding and specifying human preferences for AI systems, with the objective of overcoming obstacles related to the problem of proper specification. This work could be supported by a dedicated research hub (see recommendation 3).

To succeed in this challenge, we will need appropriate governance, as well as financial and human resources to meet the challenges. It will also require a differentiating narrative to attract the best talent.

The narrative must be that of a safe and trustworthy AI, serving the common good. Only a project clearly serving the common good will be able to attract the best talent, including foreign talent, which will be necessary to credibly challenge existing teams. For this reason, the common good mission of the project and its independence are key.

Regarding governance, this advanced research project could take the form of a Great Challenge (Grand Défi) whose objective would be to create safe and trustworthy AI general purpose AI systems. In order to have a legal structure of its own, this Great Challenge could quickly take the form of a public interest grouping (GIP). A GIP allows public and private partners to pool resources for the implementation of missions in the public interest and would have the advantages:

- of being able to rapidly mobilize top talent within partner structures, including structures with an established academic reputation capable of attracting and retaining top researchers;
- of being able to attract new talent (including, for example, open source communities) convinced by the public interest mission, provided that it is sufficiently convincing and protected.

The success of this disruptive innovation project will depend intimately on the quality of its governance and its management team, in order to make the right choices and to attract the best talent. The selection process for the management team must therefore focus on the best candidates to lead the structure, leaving aside other criteria such as nationality. An appeal to the international community will help attract the best profiles (see recommendation 1). The project's management team must then have the necessary independence to make its own choices, without the possibility of political interference. The legal and human governance of the <u>Advanced Research and</u> <u>Invention Agency</u> (ARIA), the British equivalent of the American DARPA, can be a useful source of inspiration in this respect.

This advanced research project could initially be financed up to  $\in 100$  million by the Fund for Innovation and Industry (FII), created in 2018 and endowed with  $\in 10$  billion to serve disruptive innovation (particularly in the context of the Great Challenges). The project's governance structure will then need to have the flexibility to attract financial and human resources by whatever means necessary to pursue its mission, including public, European, philanthropic, or private funding; commercial partnerships, with Al companies or providers of key resources such as computing capacity; international research partnerships; and open source contributions. It will also be able to advise on <u>calls for projects</u> from the French National Research Agency (ANR) to serve the objective of developing safe and trustworthy Al systems.

In terms of sustainable value creation, the advanced research project will aim at creating one to two companies developing state-of-the-art safe and trustworthy general purpose AI models, competitive with structures such as OpenAI, Google DeepMind or Anthropic.

## **Recommendation 3:**

Create a global research hub on understanding and specifying human preferences for general purpose AI systems. Entrust the coordination of this hub to a recognized research institute (e.g., ENS or 3IA) and ensure dedicated funding, e.g., via Priority Research Programs and Equipment (PEPR). Understanding human preferences and specifying them correctly for general purpose AI systems is not only an important technological barrier for future AI systems, but also a key research area to protect our values and limit the risk of misspecification in advanced AI systems. Recent performance gains in OpenAI's AI systems, ChatGPT and GPT-4, have come largely from its technique of Reinforcement Learning from Human Feedback (RLHF).

However, there is currently a void in the research world on this subject. Some researchers are developing techniques for AI, especially in the most advanced AI labs (see IRL, RLHF, Constitutional AI above). Others are exploring different aspects in economics, philosophy, psychology, and other scientific fields, without the objective of developing methods that can be used and systematized by machines. France and Europe have top researchers in these different fields who, when combined, could become an international reference for this nascent research field.

This hub will have to rely on the highest level of global expertise. Its implementation could be facilitated by:

- a call for international collaboration on safe and trustworthy AI (see recommendation 1);
- partnerships with researchers and AI laboratories at the cutting edge of research, which will also help to overcome the weaknesses of the French ecosystem, for example in the field of reinforcement learning. For example: OpenAI; Center for Human-Compatible Artificial Intelligence (CHAI) at the University of California, Berkeley; Anthropic; Google Brain and DeepMind (already present in France).

This cluster could be housed in an existing emblematic research institute (e.g. ENS or 3IA) and its funding could come from a Priority Research Programs and Equipment (PEPR) dedicated to safe and trustworthy AI, in the range of 50 million to 100 million euros <sup>17</sup>.

## **Recommendation 4:**

Make safe and trustworthy AI an Important Project of Common European Interest (IPCEI) to relax state aid rules and/or one of the European Union's "flagships" endowed with approximately 1 billion euros.

The development of safe and trustworthy AI systems will have to rely on a rich ecosystem of private and research actors, which France does not have on its own. To date, AI research coordinated at the European level is centered around a Network of AI Excellence Centres, which nevertheless lacks resources. (see Appendix 4).

Two tools would support ambitious European cooperation on a safe and trustworthy AI.

The European Union's "flagships" attract around 1 billion euros of investment, have a duration of about 10 years and mobilize researchers, academics, industrial companies and national programs to address major scientific and technological challenges. To date, they concern batteries, graphene, the human brain, and quantum technologies. A Safe and Trustworthy Al Flagship could conduct long-term fundamental research projects, notably on safety and trustworthiness by design, on safety issues for more advanced and even general Al systems, and on understanding human preferences and their proper specification for Al systems. This could mobilize the resources and skills of the Network of Al Excellence Centres and of the <u>public-private partnership</u> <u>on Al, data and robotics</u>, which is expected to receive 2.6 trillion euros in funding by 2030 (see Appendix 4).

The Important Projects of Common European Interest (IPCEI) of the European Union are a relaxation of European rules on state aid. They allow significant public funding for transnational European projects.

<sup>&</sup>lt;sup>14</sup>The existing <u>PEPR artificial intelligence acceleration program</u> is led by the CNRS, CEA and INRIA with a budget of 73 million euros over 5 years.

## **Recommendation 5:**

Develop two benchmarks for research to measure the trustworthiness and performance of general purpose AI systems.

Benchmarks are standardized evaluations of AI systems on a given set of tasks. By setting evaluation metrics that must be surpassed, they help guide international research efforts. The CIFAR-10 image recognition benchmark, developed by the Canadian organization CIFAR, has for example guided research in this field. It also provides transparency on technical progress through standardized evaluation.

While most AI benchmarks are quickly <u>saturated</u>, given the speed of progress in the field, there are benchmarks that measure the ability of an AI system to perform a wide variety of tasks. The most iconic of these is Google's <u>BIG-Bench</u> project, developed in collaboration with OpenAI and 132 other institutions. BIG-Bench is a benchmark used to test AI models on over 200 diverse tasks "drawing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and beyond." Such benchmarks to measure the general performance of new AI systems are particularly important in a context where the most successful AI systems are suddenly and unexpectedly developing <u>capabilities</u>. For example, natural language processing systems have suddenly developed the ability to solve arithmetic problems. Monitoring the performance of AI models on a broad set of tasks is critical to anticipate potential safety issues.

Other benchmarks measure how "ethical" or "trustworthy" an AI system is. For example, "Jiminy Cricket" environments, named after the character who guides Pinocchio's moral conscience in the Walt Disney film, were created by researchers at the University of California, Berkeley to evaluate the moral behavior of AI systems in 25 adventure games. Each action the agent can take is annotated according to several aspects of its moral character. More recently, researchers from the same university created the <u>MACHIAVELLI</u> benchmark.

Developing a European benchmark to evaluate the general performance of AI systems and a benchmark for the "trustworthiness" of an AI system would be an efficient way to monitor the progress of the best performing models in order to prevent certain risks, as well as to channel European, or even international, research towards precise safety and trustworthiness objectives. More specifically, the benchmark for general performance could be used to test whether an AI system should be considered "general purpose" in the context of European AI regulation (see recommendation 7) and the benchmark for "trustworthiness" could guide the R&D work of French initiatives (see recommendations 2 and 3).

This mission would have an estimated budget of 1 million euros.

## **Recommendation 6:**

Create a talent pool in AI safety by making public funding for AI training programmes conditional on those programs including a module on AI safety and trustworthiness.

The European Union will not succeed in developing safe and trustworthy AI leadership without talent to nurture cutting edge research and to carry a culture of safety and trustworthiness throughout the ecosystem (companies, startups, etc).

Within the framework of the second edition of France's national AI strategy, 50% of the resources, i.e. more than 750 million euros, have been dedicated to training programs.

A discipline of machine learning engineering and AI safety could therefore quickly be constituted on a national scale by making public funding of AI training courses conditional on the integration of a module on AI safety and trustworthiness. A research or higher education institute (Inria, CNRS, ENS or 3IA for example), in close collaboration with existing organizations (see box), could be in charge of developing pedagogical resources to facilitate the rapid integration of such a module and ensure the management of a community of talent in AI safety and trustworthiness.

# The EffiSciences organization to help build a community of talent in AI safety and trustworthiness

In France, the <u>EffiSciences</u> organization, founded in the Écoles Normales Supérieures (ENS), has set itself the goal of "focusing on the most pressing issues of the 21st century". In this context, it has already established several training courses on AI safety for technical profiles. In particular, it organizes conferences, research weekends on AI safety, hackathons and ten-day training bootcamps, where the most motivated students are directly involved in projects supervised by researchers specializing in these issues. Several of these students have gone on to pursue careers in AI safety at the highest level, internationally, due to a current lack of opportunities in France.

In addition, it would be appropriate to conduct a national communication campaign around AI, with a module on trustworthy AI. There are already several online AI literacy courses (Destination AI, IA Pour Tous, ClassCode IAI): they could be updated with a module on trustworthy AI and disseminated via a national online communication campaign, funded by the National Strategy for AI (SNIA).

## **Objective 2:**

Define a European regulatory framework for the safety and trustworthiness of general purpose AI and promote its adoption worldwide

## **Recommendation** 7:

Implement France's proposal to include general purpose AI systems in European AI regulation and promote its adoption worldwide via the EU-US Trade and Technology Council (TTC) and the G20.

The European Union is about to adopt the first cross-sectoral AI regulation in the world, and thus set the framework for safe and trustworthy AI (see above). However, the regulation of general purpose AI systems poses a real difficulty: the AI Act was designed to regulate AI systems according to their use case, and was not designed for general purpose AI.

Given the wide-scale safety issues that general purpose AI systems could quickly raise, it is imperative that they be subject to safety and trustworthiness requirements by design, at the time of their conception and regardless of the use case planned thereafter. This implies systematically subjecting all general purpose AI systems to the regulatory obligations of the AI Act.

France has been a forerunner on the regulation of general purpose AI systems, introducing the term in the AI Act during its presidency of the Council of the EU. It must now ensure that this term properly covers the risks associated with general purpose AI systems with three conditions:

All Al systems that are likely to be general purpose Al systems should be subjected to generality tests, allowing an initial assessment of the tasks an Al system is capable of, in order to determine whether the Al model is "general purpose" and to identify possible "high-risk" use cases. These may often be beyond the imagination and intent of their designers. The notified body, which could be the CNIL in France (the French data protection authority), must also be able to impose these generality tests if the developer of the system does not perform them spontaneously. These generality tests could be based on benchmarks aimed at evaluating the general performance of state-ofthe-art AI systems (see recommendation 5).

- General purpose AI systems should be subject to the obligations associated with "high risk" AI systems detailed in the AI Act (see Appendix 5), to the extent possible. In order to simplify these obligations, it is important that developers of general purpose AI systems have the option of relying on a generic, i.e. use case agnostic, risk management system and a generic conformity assessment procedure dedicated to general purpose AI systems.
  - If the developer of a general purpose AI system becomes aware of a "high risk" use case after the system is released on the market, this use case must be reported to the appropriate regulatory authority and the AI system subject to the associated obligations. In the case of dangerous or inappropriate use by a third party, the developer of the general purpose AI system must also take appropriate measures to limit the risk: requesting a change of use, correcting the problem, restricting or withdrawing access.

To foster innovation, it is important that the term "general purpose AI system" and the associated obligations only apply to AI systems that are truly capable of performing a wide variety of tasks. Today, this refers to a very limited number of large generative and reinforcement learning AI models developed by the largest AI companies and labs. AI systems capable of specific tasks but useful for a large number of use cases, such as speech recognition, should not be subject to these same obligations.

Standardization work, including the two working groups that France is leading on defining trust attributes and an AI risk list, will be able to bring this vision to the technical level, ensuring that each is appropriate for general purpose AI systems. This regulatory framework is also an essential tool to support the competitiveness of safe and trustworthy AI models. These models are naturally disadvantaged in a competitive environment, as they are more expensive and slower to develop, and require additional innovation.

In order to ensure a transatlantic level playing-field, a shared model of trustworthy AI should be one of the main objectives of the EU-US Trade and Technology Council (TTC). Equivalencies between European standards and those developed by NIST in the US can be established in the joint EU-NIST roadmap proposed at the December 2022 EU-US TTC.

On a global scale, France and the European Union could bring about, via the G20, a political agreement on the prohibition of AI systems that are not aligned with human interests and the common good. Once the necessary technical clarifications have been made concerning the definition of AI systems aligned and non-aligned with human interests, this could take the form of an international treaty.

## - Recommendation 8:

Entrust the future French AI regulator with a pilot experiment or a mock run of the audit process included in the EU's AI Act, in order to support the upskilling of France's audit ecosystem (companies, auditors, regulator).

The near-continuous learning and adaptation of some AI systems, especially those based on machine learning, requires a (near) continuous monitoring and auditing process, throughout the AI system's life cycle, unlike other cases of product certification or financial audits.

In this context, the future French AI regulator will have the task of structuring

and supervising the audit ecosystem, as a notifying authority under the AI Act. Audits and compliance checks can be carried out by specialized startups, auditing companies, and other organizations that have been "notified" by the regulator to be authorized to carry out this work (see Appendix 5).

Nevertheless, the French and European audit ecosystem is still far from being able to audit AI systems efficiently and requires significant efforts to invest in upskilling, new tools and new processes.

- Audit processes, whether internal or external, are far from being standardized and systematic. However, this framework is necessary to enable auditors and audited companies to move forward. One of the CEN-CLC/ JTC 21 working groups is working on the standardization of IA auditing competencies and processes.
- An ecosystem of auditors and certification bodies will have to be structured in order to bring out the leading technological players. Thus, auditors (Mazars, PwC, etc.) will have to develop their skills to understand the Al systems they will have to audit, and even develop tools to audit Al systems natively and frequently. New structures and specialized startups may also emerge to serve this market.
- Companies developing or operating AI systems will themselves have to increase their competencies and adapt their processes, especially in terms of system governance and ML Ops<sup>18</sup>. In order to facilitate the audit, these companies will be able to develop ML Ops organized with a common core, harmonizing the systems and thus simplifying the audit, to develop and trace supervision tools and processes and to have a better management of the AI information and risks between different professions: the company's manager, the compliance manager, the line manager, the quality engineer, the developer and the machine learning engineer, etc. They will also be able to adapt their SOC (security operations center) to monitor the risks linked to AI systems. These platforms, now used for the detection, analysis and remediation of cybersecurity incidents, could also track AI safety incidents.

Other developments seem desirable in the longer term to accompany the deployment of AI safety in companies:

- Insurance companies will certainly have <u>an important role</u> to play in making the audit chain accountable, provided they are able to interpret all of this continuous audit information in order to refine their risk models. Just as an insurer agrees to cover claims resulting from an airplane accident, at a cost and provided that the airplane is subject to numerous standards and audits; an insurer will be able to cover potential claims resulting from the malfunction of an AI system (which may also be embedded, for example, in an airplane), provided that the AI system is also subject to numerous standards and audits. Insurance creates strong economic incentives to measure risk and track it over time, as well as to develop the tools to do so. Cyber insurance, for example, requires specific tools and skills the same will be true for AI.
- Standardization bodies will also have to adapt to this logic of continuous control, by accelerating the development of standards capable of being read by machines (<u>SMART standards</u>). This will require significant financial, technological and upskilling investments. SMART standards also represent an important sovereignty issue: by favoring standards proposed in this format, thanks to their benefits for the company; and by reducing the margin of interpretation of standards compared to human interpretation.
- The regulator in charge of supervising the AI ecosystem will also need to have the necessary expertise. Whether this role is assigned to the CNIL or to another existing institution, the necessary upskilling will require a significant financial investment on the part of the State, and the recruitment of particularly competent personnel. France's cybersecurity authority, ANSSI, already has certain key human resources, and could provide support for certain technical aspects, or even envisage shared human resources.

<sup>&</sup>lt;sup>18</sup> ML Ops is a set of practices that aims to deploy and maintain machine learning models in production reliably and efficiently.

.

Finally, it would be useful for France's data protection agency, the CNIL, to clarify certain regulatory inconsistencies, prohibiting on the one hand the monitoring of personal data under the GDPR yet on the other soon requiring the monitoring and auditing of certain processes that will be difficult to audit without having access to these same personal data. This is for example the case of the monitoring of bias and discrimination by Al systems. In a March 2020 <u>report</u>, Institut Montaigne had proposed, for example, to adopt an active fairness approach authorizing the use of sensitive variables for the strict purpose of measuring biases and evaluating algorithms.

In this context, a first mock audit ahead of the entry into force of the European AI Act, i.e. without any sanctions for companies (reputational or otherwise), will ensure a leveling up of the auditing standards, encouraging auditors to conduct (brutally) honest assessments and companies to upgrade their standards accordingly.

An AI audit pilot experiment could also be conducted by an auditor and/ or the future French AI regulator and one or more volunteer companies between now and the entry into force of the AI Act (expected in 2025) as part of France's National Strategy for Artificial Intelligence. It would be accompanied by feedback aimed at illustrating for companies the concrete steps involved in implementing an audit process, and identifying for the public authorities the investment and guidance required to help companies, auditors and the national AI regulator adapt to the new regulatory requirements. To date, some pilot experiments of this type have been carried out, mainly by American tech players, for example the <u>Open Loop</u> initiative carried out by Meta in Estonia.

## **Recommendation 9:**

Develop a regulatory sandbox within the future French AI regulator and in close collaboration with evaluation actors such as France's National Metrology and Testing Laboratory (LNE), in order to test the conformity of new AI and general purpose AI systems before their market release.

The concept of an AI regulatory "sandbox" is provided for in the AI Act. It is a tool to facilitate the development, testing and validation of innovative AI systems before they are brought to market. A <u>pilot</u> regulatory sandbox was launched in Spain in June 2022, nonetheless with little technical means and room for maneuver.

A French regulatory sandbox, based on the model proposed by the OECD, could be entrusted to the future French AI regulator, in close collaboration with the French actors in charge of testing and experimenting AI systems (Testing and Experimentation Facilities, or TEFs), in particular France's National Metrology and Testing Laboratory (LNE). This would not only be a space for advice on how to bring an AI system into compliance with European regulations, but an environment for practically testing the degree of conformity of new AI systems with limited legal consequences. This sandbox could be based on existing tools such as LNE's Laboratory for the Evaluation of Artificial Intelligence (LEIA).

For this to happen, the regulatory sandbox concept incorporated in the Al Act must support a broader, operational vision of the sandbox, which is not limited to regulatory consultancy.

## **Recommendation 10:**

Entrust the future French AI regulator with the creation of a database documenting AI safety incidents.

The research and development of safe and trustworthy AI depends intimately on our knowledge of AI failures and incidents: when and how they occur. In many other technology areas, sharing incident reports contributes to a common knowledge base, helping industry and government track risks and understand their causes. Examples include the European database of civil aviation safety incidents (ECCAIRS) and the Information Sharing and Analysis Centers (ISACs), public-private platforms developed in the United States to share cyber incidents by sector.

#### The European database of civil aviation safety incidents

In aviation, all persons active in the field of civil aviation are <u>obliged</u> to report incidents compiled by type in the implementing regulation (EU)2015/101. If an incident is of interest but not listed, a "voluntary report" can also be sent to the authority and treated with equal attention, as the fact that events outside the list occur is in itself interesting data.

The information must be transmitted to the Authority in a format compatible with the ECCAIRS software (European Coordination Centre for Accident and Incident Reporting Systems) and the ADREP taxonomy (Accident/Incident Data Reporting) developed by the International Civil Aviation Organization (ICAO).

The Incident Reporting Guide states that incident analysis is an essential step in the event handling process. It consists of a factual description of the reported event and an interpretation of the facts. In any case, it must be proportionate to the level of risk associated with the event, ranging

from a simple assessment and closure to an in-depth analysis, the first elements of which must be transmitted to the Authority within 30 days. Associated with this document, the corrective or preventive measures that have been adopted as a result of the reported event must be attached.

In AI, some databases of safety incidents exist: the <u>Artificial Intelligence Incident Database</u> for example. In this context, different ways to encourage companies to share details of AI accidents should be explored. For example, by implementing confidentiality protections for commercially sensitive information, developing common standards for incident reporting, or mandating disclosure of certain types of incidents as provided for in the AI Act.

The creation and management of this database could initially be entrusted to the future French regulator of AI, in close collaboration with the LNE and the CEA, in order to include incidents observed in the context of the Testing and Experimentation Facilities (TEFs) and a possible "regulatory sandbox" (see recommendation 9). It could then be extended to the European level, either by the European Commission's Joint Research Centre, like the ECCAIRS database for aviation, or in the framework of the data spaces introduced in the European Union's Data Governance Act (DGA).

## Appendix 1

AI timelines: top AI researchers give 50% chance of developing human-level AI systems by 2059. So far, they have vastly underestimated the rate of development.

Top AI researchers estimate that there is more than a 50% chance of developing high-level machine intelligence, i.e. when unaided machines can accomplish every task better and more cheaply than human workers, by 2059. In the past, AI advances have far outpaced the predictions of industry experts.

Three surveys were conducted in 2016, 2019, and 2022 of AI researchers who have published in one of the two major conferences in the field-the Conference on Neural Information Processing Systems (NeurIPS) and the International Conference on Machine Learning (ICML).

The aggregate estimate of the 738 respondents to the 2022 survey was that we had a greater than 50% chance of developing AI systems by 2059 that could perform almost all tasks (>90%) better than the median human (compared to 2060 and 2061 in the 2019 and 2016 surveys, respectively). There are clear differences in opinion, however.

Even more alarming: within these surveys, when the question is asked about the likelihood that future advances in AI will be extremely harmful (to the point of posing a risk of permanent and severe impact on the human species, or even extinction), the median of researchers' answers stands at 5%.



#### Source: 2019 Survey

In the past, AI advances have far outpaced the predictions of industry experts. The 2016 <u>survey</u> compares the dates researchers estimated for some major advances with the dates of their first achievements. All the advances made to date have come earlier than the researchers' estimates: winning at Poker, and Atari games in particular, but we could also mention <u>Go</u> or <u>Star-Craft</u> games.

#### Prove math theorems Write NYT bestseller Perform well in Putnam Competition Output virtual world equations Compose Top-40 song Assemble LEGO Explain moves in computer game One-shot learning Group unseen objects Write Python code Win at Atari Assemble IKEA Autonomous vehicles drive best Win hacking competition Write history essay Transcribe human speech Beat human game speed runs Text to speech voice actor Fold laundry Human-level translation Find software vulnerability Phone banking services Win at StarCraft Win at Angry Birds Win World Series of Poker 2020 2040 2060 2080 Survey year Year 2016 : when milestone would be feasible

2019 : when milestone would be achieved and made public

## Appendix 2

## Overview of AI regulation in the world

Many national and international organizations have proposed broad principles for ethical and trustworthy AI, however, without specifying the specific attributes that these imply and without these principles being legally binding. The main ones are those of <u>UNESCO</u> and <u>the OECD</u>. These then served as the basis for the adoption of the <u>G20 AI Principles</u> in June 2019, and the launch of the Global Partnership on Artificial Intelligence (<u>GPAI</u>) in 2020, led by France and Canada within the G7.

To date, only the European Union is developing regulatory and legal frameworks that are intended to apply to all AI systems, with its draft "AI Act" and <u>Directive</u> on civil liability for AI.

In China, AI regulation is moving forward rapidly, focused for now on recommendation algorithms. In March 2022, the country adopted the "Internet Information Service Algorithm Recommendation Management Regulations" proposed by the Cyberspace Administration of China (CAC), requiring companies using recommending algorithms to inform their users and give them the option of no longer being targeted. And in January 2023, the CCA regulated "deep synthesis" technology, i.e. Al models capable of generating text, images, audio, video. In September 2021, the Chinese Ministry of Science and Technology had already published its "Ethical Standards for Next Generation Artificial Intelligence", proposing ethical principles in line with the UNESCO and OECD principles.

In the United States, soft law has been favored for the moment, carried by the National Institute of Standards and Technology (NIST), and no legislative work has been initiated. Nevertheless, in October 2022 the White House published its (non-binding) Blueprint for an <u>AI</u> Bill of Rights. It calls for greater transparency about how algorithms are created, greater accountability in Albased decision making, and the ability for users to complain if something goes wrong.

Milestones

## Appendix 3

#### Overview of AI standardization in the world

In Europe, the recognized European standards organizations CEN and CENELEC have already established the CEN-CENELEC 21 Joint Technical Committee "Artificial Intelligence" to support the development and adoption of European standards for Al<sup>19</sup>. Within this framework, they are pursuing a particularly pioneering work to define a list of safe and trustworthy Al attributes and define their technical specification by 2025, before the Al Act's entry into force.

- Due to the extraterritoriality of the Al Act, European harmonized standards will have significant reach. The Al Act, which is not yet finalized, provides that compliance with the harmonized standards proposed by CEN-CENELEC will be a means for providers to demonstrate conformity of their Al systems with the requirements of the regulation.
- 2. Internationally, the CEN-CLC/JTC 21 Joint Technical Committee actively advises other AI standardization initiatives, including ISO/IEC. Thanks to the growing expertise of the CEN-CLC/JTC 21, Europe has an opportunity to bring its vision of safe and trustworthy AI to the international level. In addition, technical cooperation agreements exist specifically between ISO on the one hand, and CEN and CENELEC on the other, to coordinate their standardization work and avoid duplication. The Commission nevertheless abides by the principle of the primacy of international standards: Europe adopts international standards when they exist or are under construction, and priority is thus given to ISO's work.

#### French participation in CEN-CENELEC working groups



**At the international level**, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) are working together to develop international standards for AI under JTC1 Subcommittee <u>SC 42</u>, Artificial intelligence.

- 1. This work is well underway, with more than 15 standards already published and others still under development.
- 2. Nevertheless, no comprehensive model of safe and trustworthy AI has been proposed: to date, ISO/IEC have not defined an exhaustive list of safe and trustworthy AI attributes, nor a guide for companies to select and prioritize the standards that are relevant to their product.
- 3. The secretariat of ISO/IEC JTC 1 SC 42 is provided by the American National Standards Institute (ANSI) and works in conjunction with 35 national standards organizations, including AFNOR in France.

# In the United States, Congress mandated<sup>20</sup> the National Institute of Standards and Technology (NIST), which is part of the U.S. Department

<sup>&</sup>lt;sup>19</sup> The main lines of the European standardization strategy have been outlined in CEN-CENELEC's <u>response</u> to the European Commission's white paper on AI, in their <u>AI roadmap</u> and in the German standardization <u>roadmap</u> for artificial intelligence.

of Commerce, to develop <u>a voluntary risk management framework for</u> <u>trustworthy AI systems</u>. An <u>initial version</u> of the *framework* was released in January 2023. NIST and CEN-CENELEC are working to align their standards, including via discussions enabled by the EU-US Trade and Technology Council (TTC). In addition, the Federal Trade Commission (FTC) recently presented its roadmap of AI system compliance requirements, focusing in particular on the risks of bias in credit granting.

# In China, the CCP is leading a standardization strategy launched in 2018 with the "China standards 2035" and specified in a roadmap in October 2021 and then in July 2022.

- 1. "China standards 2035" explicitly cites AI as a key area for standardization.
- 2. As early as August 2020, the PRC government standards body had released its <u>guidelines</u> for building a national next-generation artificial intelligence standards system. This sets a goal of implementing an AI standards system and an AI standards testing and verification platform as early as 2023. The Artificial Intelligence Subcommittee of China's National Information Technology Standardization Technical Committee (SAC/TC 28/SC 42) is in charge of this standardization work, with the <u>first</u> <u>proposed</u> standards published in July 2021.
- 3. In addition, the China Academy of Information and Communications Technology (CAICT), an influential think tank under China's Ministry of Science and Technology, is moving forward rapidly with an approach focused on developing tools to measure and test the robustness, reliability, and controllability of AI systems. The <u>white paper</u> on Trustworthy AI published by CAICT in July 2021 sets out principles that are close to those proposed by the EU and US for Trustworthy AI. CAICT works with the China AI Industry Alliance, a Chinese government-sponsored industry body, to test and certify AI systems. In November 2021, it issued its first round of trustworthy AI certifications for facial recognition systems.

Note, however, that China's Ministry of Science and Technology has not yet itself issued any documents on safe and trustworthy AI, which still leaves doubts concerning the political weight of these initiatives in the near term.

## Appendix 4

## Overview of safe and trustworthy AI research in the world

**In France**, the National AI Research Program, the research component of the national AI strategy planned for 2018-2022, has structured French AI research efforts around a network of four interdisciplinary AI institutes (3IA) and nearly 190 research chairs, including several dozen that address issues relevant to the development of safe and trustworthy AI, ranging from understanding the behavior of existing AI models, often referred to as "black boxes," to more applicative issues of certifiability.

In particular, one of the 3IA institutes, the Artificial and natural intelligence Toulouse Institute (ANITI), is specifically interested in trustworthy AI and is structured around three research programs on acceptable AI, certifiable AI and collaborative AI. The research priorities of these institutes have so far been structured by application use cases (health, environment, transportation, etc.), rather than by cross-cutting technologies or infrastructures.

France also has players capable of becoming leaders in safe and trustworthy AI R&D: industrial experts in critical systems in various sectors (aeronautics, defense, transportation and automotive, health, as well as insurance and financial services), who are at the forefront of AI innovation and who are the first to be concerned by safe and trustworthy AI.

<sup>&</sup>lt;sup>20</sup> via the National Defense Authorization Act of 2021.

More generally, France has significant assets to create or attract the best AI talent and research laboratories:

- 1. An excellent training system for AI engineers and researchers, which could very simply train safe and trustworthy AI (in 2019, France trained 6% of the best AI researchers);
- 2. An excellent AI research ecosystem (in 2019, France was home to 8% of the top AI researchers), which has already attracted many leading research labs (Google DeepMind and Meta AI Research for example);
- 3. Generous tax benefits for innovative companies: the Research Tax Credit allows a tax reduction of 30% up to 100 million euros of R&D and 5% beyond.

#### Where the best AI researchers were trained; and where they work today (<u>Macro Polo</u>)



**In Europe**, AI research is centered around a Network of AI Excellence Centres, composed of European research institutes. This network is developing 4 research projects, including two on safe and trustworthy AI: the <u>TAILOR</u> project on the foundations of trustworthy AI, and the <u>HumaneAI-Net</u> project on robust AI systems capable of understanding humans.

- While each brings together more than 50 partners, they are likely to be under-resourced. Funded with 12 million euros each over 4 years by the Horizon 2020 funding program, these efforts will continue under Horizon Europe, which funds European research over the period 2021-2027.
- In addition, the European Commission's <u>AI Coordination Plan</u> also calls for a <u>public-private AI</u>, data and robotics partnership, which is to ensure European sovereignty in the development and deployment of trustworthy, secure and robust AI, data and robotics. 2.6 trillion euros will be spent on this by 2030, including 1.3 trillion euros of funding from the European Commission.

North America, particularly the United States, is also at the forefront of most safe and trustworthy AI research topics, having a significant advantage in AI research more generally. The United States stands out for its community of AI safety researchers working on goal specification and alignment issues, which are more relevant to reinforcement learning systems and advanced or even general AI research.

 In 2021, North America accounted for 75% of publications at the FAccT (fairness, accountability and transparency) conference, compared to 17% for Europe and Central Asia and less than 5% for Asia Pacific<sup>21</sup>. The issue of robustness is of interest to defense and intelligence players.

- DARPA, the Department of Defense's applied research agency, known for pioneering innovations such as GPS, the Internet, and major developments in artificial intelligence, is conducting a project on robustness via hybrid AI (Assured Neuro Symbolic Learning and Reasoning - ANSR<sup>22</sup>), and another on deliberate
- IARPA, the equivalent of DARPA for intelligence actors, also runs two programs on AI security: Secure, Assured, Intelligent Learning Systems (SAILS) and Trojans in Artificial Intelligence (TrojAI).

The interpretability and explainability of AI systems is also an important research area for these actors.

- 1. DARPA is leading a project on explainability: Explainable Artificial Intelligence - XAI.
- 2. The U.S. National Science Foundation (NSF) and Amazon are collaborating on AI fairness, with research topics that include transparency, explainability, accountability, bias, equity, and inclusiveness.
- 3. Safe and explainable AI research company Anthropic, co-founded in 2021 by OpenAI's former VP of Research, attracted over \$700M in investment in less than a year.

Finally, some of the best AI researchers and laboratories are devoted to specification and alignment issues, especially of AI systems based on reinforcement learning, and anticipating increasingly general AI systems. In particular, we can mention:

- 1. the Center for Human-Compatible Artificial Intelligence at UC Berkeley, headed by Stuart Russell, one of the pioneers of artificial intelligence;
- 2. private American players at the forefront of artificial intelligence research: Google DeepMind and OpenAI, which have entire teams dedicated to safe AI, particularly on specification and alignment issues.

<sup>22</sup> US National AI R&D Strategic Plan 2019 update

<sup>21</sup> 2022 AI Index

 More recently, the American National Science Foundation announced a \$20M funding for a research program in AI safety ("<u>Safe Learning-Enabled Systems program</u>").

In China, while the Chinese model has traditionally been thought to be careless about safety and trust, the country's ambitions in this area and its long-term vision should not be underestimated.

Several academic and private players have been investing in the topic since the Xiangshan conference in November 2017, during which researcher He Jifeng (何积丰), introduced the concept of trustworthy Al. Among Chinese companies, JD, Tencent, and Megvii have all developed trustworthy Al initiatives. By January 2020 Megvii had established its Artificial Intelligence Governance Research Institute, and in April 2020 JD's research institute confirmed that trustworthy Al was becoming one of its main research focuses.

Moreover, China is sensitive to the issue of an integrated research strategy on trustworthy AI, incorporating general AI perspectives. In its <u>white</u> <u>paper</u> on trustworthy AI, CAICT highlights two avenues for further development of Chinese trustworthy AI research:

- 1. Develop an "integrated" AI trustworthy research agenda, to avoid working in silos and allow different AI trustworthy research projects to communicate with each other and share a common framework;
- 2. provide a roadmap anticipating the emergence of general artificial intelligence (GAI), and expand trustworthy AI research to strong AI research.

More generally, the Chinese government's increasingly cautious and conservative approach to technology may also encourage safe and trustworthy Al efforts (see the "Internet Information Service Algorithm Recommendation Management Regulations" adopted by China in March 2022).

Finally, the Chinese government understands the importance of having leading R&D players to impose its standards model. The Chinese company Huawei is an example of this.

## Appendix 5

## Conformity assessment in the EU's AI Act

The AI Act imposes conformity assessment obligations on "high-risk" AI systems. These depend on harmonized standards developed by CEN-CENELEC, which will not only set a compliance framework for AI systems, but also for AI competencies and audit processes.

Al systems that create a low or minimal risk will also be able to comply with a code of conduct that the Al Act does not specifically define. For these Al systems, it is therefore important to define the label that will structure this code of conduct at the European or even international level, and that will have to be inspired by technical standards to be legitimate.

Under the AI Act's conformity assessment procedure for AI systems, providers of high-risk AI systems must first undergo **a conformity assessment proce-dure** and then **a post-market surveillance system.** 

## In the first case, the conformity assessment can be external and performed by a "notified body". This approach concerns

- 1. High-risk AI systems used as safety components of consumer products that are already subject to third-party ex ante conformity assessments
- 2. biometric identification of persons at a distance in real time and afterwards, which do not apply the harmonized standards or common specifications.

After the conformity assessment, the notified body issues a certificate of assessment of the EU technical documentation; the supplier then draws up an EU declaration of conformity, affixes the "CE" to the product, and draws up an EU declaration form. A notified body is a conformity assessment body designated by the notifying authority of the country in question. For example, for personal protective equipment (PPE), in France, the *Direction Générale du Travail* (DGT) is the notifying authority. **Concerning the AI Act, it is possible that the CNIL will have the role of notifying authority. Certification bodies (specialized startups, audit companies, and others) will have to be notified by this authority to be authorized to carry out this work.** 

In a second case, the conformity assessment may be internal and performed by the supplier itself. The vendor may work with third parties such as auditing companies. This concerns autonomous high-risk AI systems (i.e. not concerned by cases 1/ or 2/ above). In this case, the provider must comply either with harmonized standards (if they exist); or with common specifications. After the conformity assessment, the provider then draws up an EU declaration of conformity, then affixes the "CE" to the product, and then draws up an EU declaration form.

A post-market surveillance system is to be put in place when the Al system is on the market to assess continued conformity. For Al systems that have received an external conformity assessment, the Notified Body also performs periodic audits.

## Acknowledgements

The author of this note would like to thank all the members of the Institut Montaigne team that made it possible to carry out this study, in particular Tom David, Assistant Policy Officer, and Camille Le Mitouard, Project Officer, as well as all the people interviewed or consulted while conductingin this work:

- Jamal Atif, Professor and Vice President in charge of the Digital Sector, University Paris Dauphine-PSL
- Guillaume Avrin, National coordinator for artificial intelligence
- Francis Bach, Research faculty, Inria École Normale Supérieure (ENS)
- Pierre-Etienne Bardin, Chief Data Officer, La Poste
- Annabelle Blangero, Data Scientist Senior Manager, Ekimetrics
- Anne Bouverot, Chairperson of the Board at Technicolor and Chairperson of Fondation Abeona
- Raja Chatila, Professor emeritus, Sorbonne University
- **Julien Chiaroni**, Former Director of Grand Défi in Artificial Intelligence, French General Secretariat for Investment (SGPI)
- Rémy Choquet, Head of Medical Evidence and Medical Data Platforms and Data Sharing, Roche
- Marie-Pierre de Bailliencourt, Managing Director, Institut Montaigne
- Caroline de Condé, Head of Digital Standards Unit, Groupe AFNOR
- Marcin Detyniecki, Head of Research and Development & Group Chief Data Scientist, AXA
- Marko Erman, SVP, Chief Scientific Officer, Thales
- Laurent Inard, Partner and Chief R&D Officer, Mazars
- Caroline Jeanmaire, Doctoral Researcher in Artificial Intelligence Policy, Blavatnik School of Government, Oxford University
- Elliot Jones, Researcher, Ada Lovelace Institute
- **Fabien Le Voyer**, Deputy national coordinator for artificial intelligence
- **Emmanuelle Legrand**, AI Project manager (regulation), French Directorate General for Enterprise (DGE)

- Bruno Maisonnier, CEO and Founder, Another Brain
- Nicolas Marescaux, Deputy Director, Meeting Customers Needs & Innovation, MACIF
- Sébastien Meunier, Public Affairs Manager, ABB France
- Nicolas Miailhe, Co-founder and President, The Future Society
- Nicolas Moës, Director European Al Governance, The Future Society
- Louis Morilhat, AI Officer, Groupe AFNOR
- Aurélien Palix, Head of Digital Networks and Uses division, French
  Directorate General for Enterprise (DGE)
- Ludovic Peran, Product Manager for Responsible & Human-centered Al, Google Research
- Tanya Perelmuter, Co-founder and Director of Strategy and Partnerships, Fondation Abeona
- Gabriel Peyré, Professor, CNRS, DMA, École Normale Supérieure (ENS)
- Hadrien Pouget, Visiting Research Analyst, Carnegie Endowment for International Peace
- Timothée Raymond, Head of Innovation & Technology, Linedata
- Benoit Rottembourg, Head of REGALIA Algorithmic Audit & Regulation, Inria
- Gérard Roucairol, Honorary Chairman, National Academy of Technologies of France
- Stuart Russell, Professor of Computer Science, University of California, Berkeley
- **Isabelle Ryl**, Director, PRAIRIE (PaRis Artificial Intelligence Research InstitutE) Inria
- Guillaume Sylvestre, Director Digital Intelligence, ADIT
- Helen Toner, Director of Strategy and Foundational Research Grants, Center for Security and Emerging Technology (CSET)
- Fabrice Valentin, VP Head Of Advanced Analytics & Artificial Intelligence, Airbus

The opinions expressed in this report are not necessarily those of the abovementioned persons or the institutions that they represent. Institut Montaigne welcomes thoughts and ideas on how to address these issues collectively and put forward recommendations which serve the public interest.

## Members

ABB France



Abbvie Accenture Accuracy Adeo ADIT Aéma Air France - KLM Air Liquide Airbus Allen & Overv Allianz Amazon Amber Capital Amundi Antidox Antin Infrastructure Partners Archery Strategy Consulting Archimed Ardian Arauus Astrazeneca August Debouzy Avril AXA **Baker & Mckenzie** Bearingpoint Bessé **BG Group BNP** Paribas Bolloré Bona Fidé Bouygues **Brousse Vergez** Brunswick Capgemini **Capital Group** CAREIT Carrefour Casino Chubb CIS **Cisco Systems France Clifford Chance** Club Top 20 CMA CGM

CNP Assurances Cohen Amir-aslani **Compagnie Plastic** Omnium Conseil supérieur du notariat Crédit Agricole D'angelin &Co.Ltd Dassault Systèmes De Pardieu Brocas Maffei DIOT SIACI Doctolib ECL Group Edenred EDF EDHEC Business School Eais **Ekimetrics France** Enedis Engie EOT ESL & Network **Ethique & Développement** Eurogroup Consulting FGS Global Europe Fives Getlink **Gide Loyrette Nouel** Google Groupama Groupe Bel Groupe M6 Groupe Orange Hameur Et Cie Henner **Hitachi Energy France HSBC Continental Europe** IBM France IFPASS Inkarn Institut Mérieux International SOS Interparfums Intuitive Surgical Ionis Éducation Group iQo ISRP Jeantet Associés

Jolt Capital **Kantar Public** Katalyse Kearnev **Kedge Business School** KKR KPMG S.A. La Banque Postale La Compagnie Fruitière Linedata Services Lloyds Europe L'Oréal Loxam LVMH - Moët-Hennessy **Louis Vuitton** M.Charraire MACSE MAIF Malakoff Humanis Mazars **Média-Participations** Mediobanca Mercer Meridiam Michelin **MicroPort CRM** Microsoft France Mitsubishi France S.A.S Moelis & Company **Moody's France** Morgan Stanley Natixis Natural Grass Nestlé OCIRP ODDO BHF **Oliver Wyman Ondra Partners** onepoint Onet Optigestion Orano **Ortec Group PAI Partners Pelham Media** Pergamon Prodware **PwC France & Maghreb** 

Raise RATP **RELX Group** Renault Rexel **Ricol Lasteyrie** Rivolier Roche **Rokos Capital** Management **Roland Berger Rothschild & Co** RTE Safran Sanofi SAP France Schneider Electric Servier SGS SIER Constructeur SNCF SNCF Réseau SNEF Sodexo SPVIE SUEZ Taste **Tecnet Participations SARL** Teneo The Boston Consulting Group Tilder Tofane **TotalEnergies UBS France** Unibail-Rodamco Veolia Verlingue VINCI Vivendi Wakam Wavestone Wendel White & Case Willis Towers Watson France Zurich

Institut Montaigne 59 rue La Boétie, 75008 Paris Phone +33 (0)1 53 89 05 60 institutmontaigne.org

Printed in France Legal deposit: April 2023 ISSN 1771-6756 Al, and with it our society, is at a historical turning point. We are now developing "general purpose" Al systems like ChatGPT, capable of performing a large number of tasks. This could quickly become a decisive competitive advantage for companies and countries alike.

However, these systems represent a major and growing security challenge. Not only because they can be used by malicious actors, but also because the statistical nature of today's AI systems poses unprecedented safety risks, which are now some of the most important technological barriers in the field.

This challenge also represents a unique opportunity for France to position itself as a leader in safe and trustworthy AI, by attracting some of the best AI talent, for whom safety is becoming a major concern and yet not properly addressed by their current employers. It has world-class researchers in mathematics and AI, as well as cutting edge expertise in systems and software engineering for safety. Thanks to the powerful computers of the French National Centre for Scientific Research (CNRS), it is also one of the only European countries able to develop large, general purpose AI models.

To seize this opportunity, France must give itself the means to do so, with a disruptive innovation project and a fundamental research cluster dedicated to the development of safe and trustworthy general purpose AI systems. It must also ensure that the cutting edge but potentially dangerous AI systems developed today by American and Chinese actors are subject to future European regulations, which are likely to define the international requirements for AI safety and trustworthiness.

> 10 € ISSN : 1771-6756 NAC2304-01