INSTITUT Montaigne

institutmontaigne.org
NAC2304-01

POLICY PAPER - April 2023

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

Although artificial intelligence (AI) 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» AI systems like ChatGPT are now capable of performing an increasingly large number of tasks and could rapidly accelerate technological change.** AI 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 AI 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 quickly be misused for malicious purposes. AI 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 AI. Europe has accumulated a delay that is difficult to make up. **Focusing on safe and trustworthy AI now constitutes Europe's best differentiation strategy to position itself as a key player in AI. 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) 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 AI safety by making public funding for AI training programs conditional on those programs including a module on AI 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 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.

Recommendation 10:

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