CENTRE OF ADVANCED STUDIES FOR ARTIFICIAL INTELLIGENCE

CSAIA
CENTRO DI STUDI AVANZATI
PER L'INTELLIGENZA ARTIFICIALE

CSAIA Declaration of Intent

AI at service of Rule of Law

Introduction

Considering the nature, goals, and objectives of CSAIA, already expressed in its charter,

Considering that AI embraces both challenges and opportunities at the global level,

Considering the critical historical moment for multilateral International relationships,

Considering the need to address the challenges and opportunities provided by AI through European and International cooperation that goes beyond physical and intellectual frontiers with an interdisciplinary and cross-disciplinary vision,

Considering that scientific dialogue is an essential tool for ensuring informed and forward-looking choices,

Considering that CSAIA is a forum for international scientific discussion to identify and implement best practices and best solutions in AI,

Convinced that the development of AI systems must remain in the service of mankind and the Rule of Law and contribute to the preservation of Peace,

In compliance with the highest European and International standards and applicable regulations,

Addressing National, European and International Institutions to begin a constructive and operational dialogue based on common principles

With the aim of facilitating the resolution of di culties through interdisciplinary, cross-cutting and international scientific treatment of complex issues related to AI developments and applications,

With the aim of reinforcing the awareness of all actors,

CSAIA proposes the following Declaration of Intent for the adherence of Institutions willing to participate in this constructive dialogue and approach.

Program Points

1. Design and privacy and security by design, intellectual property rights:

- the design of an AI system may follow, in compliance with internationally recognized privacy and security rules, globally recognized tools and best practices using a structured methodology called pipeline, that provides for the phases of data collection, data processing, data refinement, modeling, model evaluation and model publication according to the standards of the CRISP-ML (Q) methodology. For each identified and developed use case developed, a pipeline-type architecture may be contemplated, ensuring that the most suitable technologies are used (generative AI, symbolic AI, and traditional computing, or a hybrid of the three).

- with regard to copyright aspects, an AI model shifts the value of information from the data to the model capable of extracting tangible information elements from the data, of their interpretation, processing, synthesis, prediction, and construction of new concepts that become the property of the institution that trained the model and remain under its responsibility

- AI for Security. Without prejudice to the objective of protecting the country's physical, logical and network security, AI can truly be a valuable support for those responsible for protecting and ensuring the "always on" status of critical industrial infrastructure, public administration, healthcare, telecoms and IT infrastructure, as well as infrastructure related to national security and space. The increase in malicious actors, groups sponsored by hostile states, and terrorism mixed with crime require us to look to AI as a possible and realistic technological resource for defence and counteraction, even in the presence of malicious AI used to foment a hybrid technological conflict. This hybrid conflict manifests itself through the propagation of fake news, deep fakes, fake social media profiles, honey pots, fake blogs, ransomware, DDOS, trolls and influence on the dominant thinking of a country. This is why CSAIA has decided to implement this line of speculation and study in the service of the community.

2. Human in the loop and Ethical compass:

- the design of AI systems must always be inspired by the concept of cohabitation of technology with human capability,with the goal of extending its abilities but never replacing it

- fine-tuning of generative AI models should always take place under the close supervision of human involvement and never be allowed to automatically make decisions that could have potentially severe consequences on and negatively impact the individual

- the AI system should always be designed to be modifiable via human involvement in order to respect the right to oblivion (aka right to be forgotten)

- AI-based systems must comply with the globally recognized ethical principles of transparency, privacy, fairness, explainability, safety, robustness, accountability, and human control

3. Governance and Segregation:

- model development, data preparation, management and deployment, and performance monitoring are activities that a governance platform must ensure through machine learning operations and governance tools for the development and deployment phases

- in the event of exclusive use of any AI model (including generative AI), the adaptation of the model to the context should be operated with machine learning using lawful and unbiased input data avoiding external contamination and ensure the involvement of synthetic data or other privacy engineering concepts in order to minimize the phenomenon of hallucinations, overfitting and cognitive bias

- international collaboration is needed for initiatives aimed at fostering collaboration among legal scholars, AI researchers, policy makers and practitioners to create a comprehensive framework to address the legal issues of generative AI

4. Machine Unlearning (MU):

- the AI system should always be designed to be modifiable via human involvement in order to respect the right to be forgotten and other relevant individual rights Objectives include:

- compliance with privacy legislations such as the EU GDPR (which also includes the right to be forgotten) would enable removal of personal data from trained models, safeguarding the privacy of individuals

- increased security against “data poisoning” attacks by removing so-called toxic data aimed at manipulating a model’s behavior

- improved adaptability of models to remain relevant even when data distributions change over time.

- regulatory compliance with changing laws and regulations

5. Regulatory Compliance:

AI-based systems must be secure and comply with existing fundamental rights and values in line with applicable privacy and other relevant legislation. .

In the EU, the GDPR defines the requirement for lawful use of personal data by AI models. AI developers must via embedded technology ensure that the required permissions and authorizations, such as explicit consent, is provided voluntarily, and that it is specific, informed, and unambiguous. In some cases, AI may process personal data on the basis of other lawful bases such as “legitimate interest.”

6. Contamination:

Knowledge of the developed designs and their repurposing in intended cases must be fostered in order to explain how knowledge-sharing can in practice take place and what benefits it can bring.

7. Pedagogical context and AI: Emendative and emancipatory perspectives in education

7.1. The pedagogical foundation: emancipatory and corrective perspectives

In the contemporary socio-educational context, AI is not just a tool, but an environment that requires a redefinition of critical thinking. We confirm the centrality of an emendative pedagogy, which values error (including algorithmic error) as an opportunity for learning and refinement, and of an emancipatory pedagogy, which promotes human autonomy against all forms of passive technological dependence. In this vision, education is the first ethical safeguard: it serves to deconstruct the "black box" and rebuild transparent knowledge, overcoming the "fear of education" to embrace conscious management of complexity.

7.2. The operational gap: competence as a defense of rights

However, theoretical awareness is not enough to guarantee the rule of law within complex organizations and public administration. It is necessary to bridge the gap between ethical principles and daily operations ("Head vs. Hands gap"). The principle of "Human in the Loop" risks being emptied of meaning if the human in the loop does not possess the technical skills to exercise real control. Therefore:

- from presence to competence: human supervision of AI systems is only valid if supported by certified and up-to-date training. An untrained operator is not a guarantor, but a potential point of vulnerability for citizens' rights and data security;

- hybridization of knowledge: there is an urgent need to promote hybrid training courses, where lawyers acquire technical basics and technicians understand regulatory boundaries, creating a common language essential for compliance and governance.

7.3. From fear to mastery

To protect democratic institutions and values, training must evolve from simple "literacy" to operational mastery.

- Operational capability as a guarantee: knowing how to correctly interrogate a model, recognize a hallucination, manage the privacy of input data, and evaluate output are not just technical skills, but acts of civic and administrative responsibility.

- Adaptive reskilling: institutions are committed to overcoming the rigidity of traditional training paths by adopting agile lifelong learning models capable of keeping pace with the evolution of AI models (e.g., Agentic AI) to ensure that the defense of the rule of law is always technologically up to date.

8. Agentic IA:

The autonomy of Agentic AI poses both ethical and legal challenges: it requires robust governance, human control mechanisms, and constant alignment with current and evolving regulations. Assessment of ethical risks (autonomy vs. human control, bias and dsicrimination, transparency and explainability, privacy and oversight, accountability and responsibility) and legal risks (liability and responsibility, regulatory compliance, intellectual property protection, cybercrime) need to be constantly present.

9. AI in Finance:

AI is now an essential component of the process of designing, configuring, and delivering increasingly personalized banking and financial products and services. When a model recalibrates credit granting criteria and related pricing, prioritizes anti-money laundering alerts, estimates expected losses, or intercepts payment fraud, it affects both the technical profiles of intermediaries and the degree of customer protection and trust in the financial system. The responsible adoption of AI therefore requires informed governance choices that must be left to the strategic supervisory bodies of financial institutions, rather than to their respective IT managers.

Given AI's ability to influence the entire financial intermediation process, model-centric controls comparable to those reserved for critical functions are needed: explicit roles and responsibilities, three lines of control, traceability of decisions, auditability of the life cycle, and proportionality with respect to the potential impact on customers, markets, and systemic stability. Each model should have an “operational passport” that sets out its purpose, data, limits, metrics, controls, as well as rules on permitted autonomy and end-to-end traceability (versions, inputs, outputs, human interventions).

The use of generative models and agentic systems requires additional precautions, including the timely detection of plausible but incorrect responses, the creation of dedicated testing environments, the complete recording of actions performed, defense systems against manipulation (data poisoning, prompt injection, exfiltration), and the provision of limits on automatic action with immediate kill switch mechanisms (kill switch).

In addition, processes that may have significant economic impacts or affect customer rights and interests must be designed to allow for the rapid reversal and correction of decisions (reversibility by design), with structured incident management procedures (escalation, suspension, analysis, and remediation).

Finally, systemic risks related to concentration on a few model and cloud providers, interdependencies in the supply chain, and potential pro-cyclical effects are relevant. To this end, portability requirements such as the ability to transfer data, logs, and controls and to replace models or suppliers without loss of operational continuity, through exit plans and periodic migration tests, are essential to protect market integrity and financial stability.

10. Artificial Intelligence in Healthcare:

– Since the negotiating position of June 14, 2023, the EU Parliament has identified as high-risk AI systems those that affect the healthcare sector. In light of recent case law and in compliance with the European Regulation, the key principles that must govern the use of algorithms and AI tools in carrying out tasks of significant public interest—such as healthcare—include:

the principle of knowability, under which the data subject has the right to be informed of the existence of decision-making processes based on automated processing;

the principle of non-exclusivity of algorithmic decisions, requiring that there must always be a human intervention within the decision-making process capable of verifying, validating, or overturning the automated outcome;

and the principle of non-discrimination in algorithms, which requires that data controllers use reliable AI systems subject to periodic verification.

– Regarding the development of AI systems to be applied in healthcare, the accuracy, correctness, and updating of data are of particular importance, given the risks linked to the processing of health data initially collected for care purposes, which may later have been modified, corrected, or supplemented by healthcare personnel involved over time in the patient's care pathway.

– In line with international recommendations dating back to 2019 (OECD), it is essential in healthcare to adopt a human-centered approach at every stage of the AI system lifecycle, and to uphold the principles of fairness and non-discrimination, transparency and explainability, robustness and cybersecurity, as well as the principle of accountability for all those who, in various capacities, play an active role in the design, development, and deployment of artificial intelligence systems.

– Finally, in accordance with the recommendations of the European Data Protection Supervisor (EDPS) and the European Data Protection Board (EDPB), it is necessary in the healthcare field to maintain the central role of healthcare professionals during the training phase of algorithms, to ensure that decisions are not generated solely by machines. This process must be thoroughly documented in the Data Protection Impact Assessment (DPIA) in order to mitigate risks to the rights and freedoms of data subjects—particularly the risk of algorithmic discrimination, which could affect access to healthcare, patient cost-sharing, and even the appropriateness of diagnostic and therapeutic pathways.

Contributor:

- **Marco Braccioli** - ICSA Foundation Councillor, Member of the Executive Board of Afcea International-Rome Chapter, Member of the Technical Committee of CESMA (Centre for Military Aeronautical Studies), Rome

- **Ulrika Dellrud** – Global chief privacy and data ethics officer, Board Member and NED, Certification Director ISACA Belgium, AI Governance Advisory Board Member, IAPP, Advisor to European Data Protection Board, Member of the NY Bar, Bruxelles

- **Agostino Ghiglia** - GPDP College Member, Data Protection Authority, Rome

- **Simon Greenman** - Chief AI Officer Best Practice AI, past World Economic Forum Global AI Council, Harvard Business School Alumni Angels UK, London

- **Konstantinos Kenanidis** - Secretary General Office of the Orthodox Church in EU, Bruxelles

- **Suany Mazzitelli** - Éditeur LexisNexis, membre Observatoire de légistique, Paris

- **Francesca Medda** – Director at Institute of Finance and Technology, UCL, London

- **Pier Paolo Maria Menchetti** - EU Commission/EMA Chairman Expert Panels spinal devices in Orthopaedics, Traumatology, Rehabilitation, Rheumatology; co-Director International School on AI Technology and Law - Ettore Majorana Foundation and Centre for Scientific Culture, Erice

- **Vivaldo Moscatelli** - Ambassador EU Community EDSC project (European Digital Skill Certificate), Rome

- **Saverio Occhipinti** - General Counsel Club Deal Investors University LUISS Guido Carli ("LUISS Alumni 4 Growth"), LUISS University, Rome

- **Sun Kun OH** - past Korean delegate OECD; Emeritus Professor of Physics, Konkuk University, Seoul, South Korea

- **Anna Pouliou** - Maastricht University, Faculty of Law, European Center of Privacy and Cybersecurity; Data Protection Commission, CERN, Genève

- **Daniele Pulcini** - University for Peace - UN Mandated, Rome

- **Domenico Franco Sivilli** - General Director for Information Resources and Statistics, Council of State, Rome

- **Cinzia Turli** - Pedagogy. San Raffaele Telematic University, Rome; G. D'Annunzio University, Chieti Pescara

- **Andrea Vivoli** - General Director of Central Bank of San Marino Republic, San Marino

- **Fabrizio Zichichi** - Vice President at Ettore Majorana Foundation and Centre for Scientific Culture, Erice

Associazione CSAIA
ETS rep. n. 135104

Viale Città d'Europa, 681
00144 ROMA.
Tel. +39 335 8005403
www.centrostudicsaia.it
info@centrostudicsaia.it