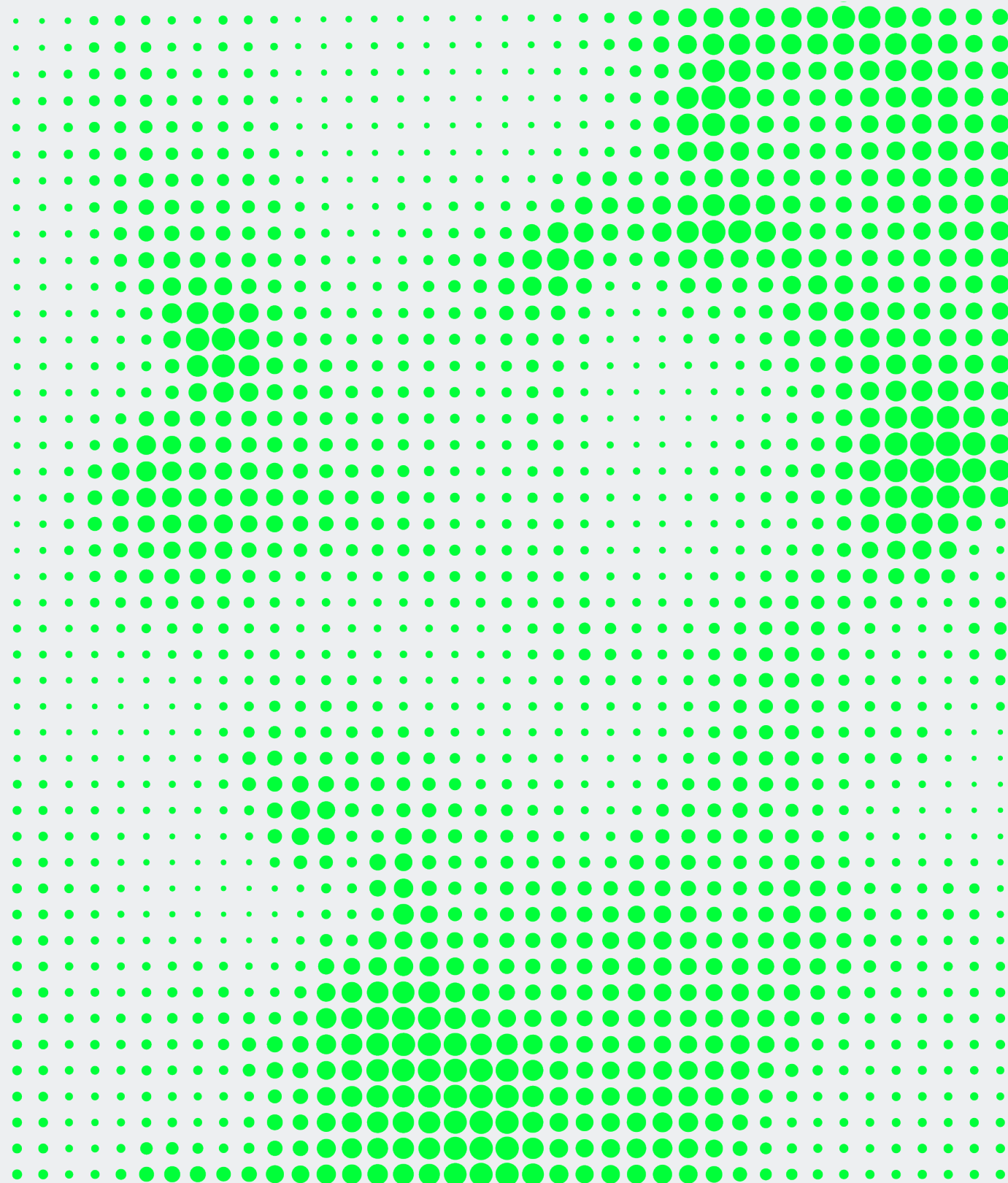




INTERNATIONAL COUNCIL ON ARTIFICIAL INTELLIGENCE

October 2026

MADRID MANIFESTO FOR RESPONSIBLE AI

Towards a democratic, innovative, inclusive, sustainable
and responsible AI that serves the public interest



COUNCIL MEMBERS

Manuel Castells

Chair. Professor of Communication Technology and Society at USC and Professor Emeritus at UC Berkeley.

Luciano Floridi

John K. Castle Professor in the Practice of Cognitive Science and Founding Director, Digital Ethics Center, Yale University; Professore di Sociologia della Cultura e della Comunicazione, Università di Bologna; expert in AI ethics.

Paul S. Adler

Harold Quinton Chair Emeritus in Business Policy and Professor Emeritus of Management and Organization, Marshall School of Business, University of Southern California.

Jeroen van den Hoven

Professor of Ethics and Technology at Delft University of Technology and Scientific Director of Delft Design for Values.

Francesca Bria

Honorary Professor at the UCL Institute for Innovation and Public Purpose, President of ART-ER (Emilia-Romagna's innovation agency) and Board Member of the European Innovation Council; expert in digital sovereignty.

Carissa Véliz

Associate Professor at the Institute for Ethics in AI, University of Oxford, and author of Privacy Is Power.

Vint Cerf

Internet pioneer and recipient of major distinctions, including the US National Medal of Technology and the Turing Award.

Erika Staël von Holstein

Chief Executive of the independent think tank Re-Imagine Europa; expert in AI and political polarisation.

Kate Crawford

Professor at USC Annenberg; Inaugural Chair in AI and Justice at the École Normale Supérieure; leading expert on the social, political and environmental implications of AI.

Niklas Lundblad. Expert in technology policy and responsible technology, and formerly a global executive at Google, with over twenty years of experience in tech policy and research.

Jerome A. Feldman. Professor Emeritus of Electrical Engineering and Computer Sciences at UC Berkeley; a leading figure in computing and cognitive science.

BACKGROUND

The International Council on Artificial Intelligence was established in June 2024 to support the Government of Spain in addressing one of the defining challenges of our time: understanding and responding to the opportunities and challenges posed by artificial intelligence, while guiding its development towards shared progress, the protection of rights and the strengthening of democracy.

Chaired by Manuel Castells and composed of internationally recognised experts, the Council brings together a broad range of disciplines and perspectives to provide independent strategic advice on the long-term implications of artificial intelligence.

The Madrid Manifesto is the outcome of the Council's work since its establishment in 2024 and embodies the collective reflections, deliberations and recommendations developed throughout its successive meetings. It brings together the positions agreed by the Council's members and expresses a shared conviction: artificial intelligence will achieve its fullest potential when societies are able to integrate it effectively in ways that benefit people, strengthen social cohesion and expand opportunities for progress. The Manifesto therefore seeks to transform the Council's accumulated knowledge into a common reference point for a democratic, innovative, inclusive, sustainable and responsible artificial intelligence.

DISCLAIMER

This Manifesto reflects the collective work and broad consensus of the members of the International Council on Artificial Intelligence. Members participate in a personal capacity and not as representatives of any government, institution or organisation. While the Manifesto expresses a shared vision and common principles agreed by the Council, it should not be understood that every member necessarily endorses each individual statement in its entirety.

Nothing in this Manifesto shall be interpreted as representing the official position of any government, institution or organisation with which a Council member may be affiliated. Furthermore, this Manifesto does not constitute or represent an official position of the Government of Spain.

INDEX

Preamble

I Human-Centric Artificial Intelligence at the Service of People and the Common Good

- Human-centric AI
- Respect and protection of human rights and civil liberties
- Protection and enhancement of human autonomy
- Equality, inclusion and non-discrimination
- Privacy and data minimisation
- Enhanced protection of minors

II Sovereign and Democratic Artificial Intelligence

- Protection and strengthening of the public sphere
- Digital sovereignty
- Cognitive autonomy
- Resilience
- Institutional, Public and Societal Capacities
- Digital infrastructures serving the public interest

III Inclusive, Innovative and Sustainable Artificial Intelligence

- Legitimate purpose
- Technological proportionality
- A broad concept of progress
- Education and training
- Enhancement of work
- Social, linguistic and cultural diversity
- Intergenerational, material and territorial sustainability

IV Responsible, Transparent and Trustworthy Artificial Intelligence

- Identity and transparency
- Trustworthiness and reliability
- Traceability, auditability and redress
- Assessment throughout the entire lifecycle
- Meaningful human control
- Justice and Accountability

PREAMBLE

Artificial Intelligence (AI) can offer genuine opportunities for progress. If designed and implemented correctly, it may accelerate scientific discovery and medical diagnosis, broaden access to knowledge and education, improve the quality of public services, and enable persons with disabilities to participate on more equal terms. These opportunities are precisely why it is worth developing AI properly.

AI is also a geopolitical force and a cross-cutting social, economic, informational and political infrastructure that is transforming every sector and sphere of activity. It is therefore essential to promote AI that is consistent with and supportive of the principles and values of democratic societies.

Power over AI is increasingly concentrated in the hands of a small number of corporate actors. Major economic and technological actors exercise de facto power over the conditions of everyday life monopolizing expertise, data and tools. Companies and platforms define conditions for access, rules of visibility, forms of interaction, and even economic opportunities. In the digital era, colonialism assumes new forms. It no longer dominates only bodies, but appropriates data, transforming personal lives into exploitable information. Freedom in the digital age is not merely a matter of interiority but also a public concern. It calls for clear rules, transparency, the possibility of recourse and proportionate limits on the use of intrusive technologies, so that technology will remain at the service of the human person.

Such a situation raises fundamental challenges for democratic governance, economic fairness and technological sovereignty. Self-regulation cannot ensure that AI develops in ways that advance the public interest. Democratic societies therefore have a legitimate role in shaping, governing and overseeing AI so that its benefits, opportunities and capabilities are shared broadly and aligned with human rights and democratic values.

The digital transformation of recent decades has taught us valuable lessons above all about how power and value tend to concentrate when a small number of actors control the infrastructure upon which others depend. AI offers the opportunity to apply these lessons and to build differently.

AI can become a means of concentrating knowledge and decision-making in a small

number of systems, or a means of enabling many people to draw on knowledge that remains in the hands of those who create it, and to develop solutions adapted to their own needs and contexts. Which path we take depends on foundations that are being laid now: how data is accessed, how creators and other contributors are recognised and rewarded, how information is trusted and how systems connect with one another. These foundations are best shaped openly, with all those who will build on them. On clear rules and trusted infrastructure accessible to all, AI can become a source of creativity and opportunity: for millions of small and medium-sized enterprises, start-ups, creators, researchers and communities across Europe and beyond that can use it to solve the concrete problems of their sectors and territories.

Our shared objective is to guide, regulate and govern AI so that the knowledge, power and benefits it generates strengthen autonomy, contribute to economic and social well-being and cohesion, and remain clearly aligned with democratic values and the broader public interest. Beyond the capabilities of AI systems, success depends on our commitment to human-centric, democratic, innovative, inclusive, sustainable and responsible AI that serves the public interest. That choice remains ours to make.

For the purposes of this Manifesto, references to AI shall be understood to include those who design, develop, deploy, implement, operate and govern AI systems.

SECTION I

HUMAN-CENTRIC ARTIFICIAL INTELLIGENCE

AT THE SERVICE OF PEOPLE AND THE COMMON GOOD

01 Human-Centric AI

AI must serve people, society and the common good. Technology should complement human intelligence, creativity, professional expertise and well-being.

02 Respect and Protection of Human Rights and Civil Liberties

Human Rights and Civil Liberties must be embedded by design and throughout the development, deployment and implementation of AI systems.

These rights and liberties, together with ethical principles, must be translated into verifiable technical and organisational requirements.

03 Protection and Enhancement of Human Autonomy

AI must respect and promote human autonomy as the capacity and freedom to think, act, decide and exercise control over one's own life.

Predictions and recommendations generated by AI systems may inform decisions that significantly affect people, but must remain distinct from them, so that responsibility for each decision rests with an identifiable person or institution able to explain it and answer for it. A clear and effective separation should be maintained between prediction, recommendation and decision.

04 Equality, Inclusion and Non-Discrimination

AI should be designed, developed, deployed and assessed with the objective of promoting equality, fostering inclusion and ensuring non-discrimination.

When appropriately designed and deployed, AI can be a valuable tool in reducing inequalities arising from a lack of resources, skills, connectivity or accessibility.

05 Privacy and Data Minimisation

Privacy is both an individual right and a prerequisite for freedom and democracy. Mass surveillance is in fundamental tension with human rights and civil liberties. Protecting privacy in the use of AI is a means of supporting a fair and balanced distribution of power.

Data minimisation is a basic principle. All data processing must pursue a legitimate objective, and only the data necessary to achieve that objective should be used.

06 Enhanced Protection of Minors

The protection of minors is not a constraint on innovation; it is the condition that legitimises innovation. Children and adolescents are rights-holders whose best interests must guide the design, development and deployment of AI. AI companions and chatbots should be designed to support the well-being, autonomy and healthy development of children and young people.

AI systems likely to reach minors should be subject to independent assessment before deployment to ensure high standards of safety, protection and accountability.

Every technology transforms the world that future generations will inherit. Our future and that of our children must come before the profits of the AI industry. The responsibility of this generation is to ensure that AI expands their opportunities through safety, age-appropriate design, digital literacy and positive educational outcomes.

SECTION II

SOVEREIGN AND DEMOCRATIC ARTIFICIAL INTELLIGENCE

07 Protection and Strengthening of the Public Sphere

AI should strengthen democracy and the public sphere by helping create the conditions for individuals to form independent opinions and participate in public life.

AI can be used to improve democratic deliberation by faithfully summarising opposing viewpoints, identifying areas of agreement and facilitating multilingual public consultations. AI can make a positive contribution to democracy. For AI to play this role, the channels through which people increasingly obtain information, including conversational AI systems, must remain plural and accountable.

08 Digital Sovereignty

Digital sovereignty combines three inseparable dimensions: democracy, security and competitiveness. It ensures citizens' informational self-determination, keeps critical digital infrastructures under democratic oversight and European law, and enables Europe to develop and govern key technological capabilities so that the benefits of AI are broadly shared across society.

Digital sovereignty is built through openness and partnership. It means ensuring that people and institutions retain meaningful control over the data, systems and services on which they depend, and that AI reflects the linguistic, cultural and territorial diversity of the societies it serves. Systems that draw on diverse languages, contexts and knowledge serve citizens better, and open standards and interoperability allow that diversity to become a shared strength.

Openness, interoperability and international cooperation are fully compatible with strategic autonomy when they safeguard the capacity of democratic societies to make their own choices and shape their own technological future.

Digital sovereignty is a necessary condition for democracy. It is the capacity of democratic societies to govern critical digital infrastructures in accordance with their laws, values and collective priorities. In the digital age, sovereignty does not mean isolation, but rather autonomy, resilience and democracy.

AI infrastructure constitutes a strategic foundation for economic prosperity, democratic resilience and social cohesion. Democratic societies should maintain robust public capabilities across the AI stack to ensure that critical technological infrastructures remain subject to democratic oversight and responsive to the public interest. Public capacity is a cornerstone of democratic sovereignty.

09 Cognitive Autonomy

Cognitive autonomy is the capacity of individuals to form their own beliefs and make their own decisions, preserving free will from undue interference derived from technology or digital systems.

Rooted in freedom of thought and the right to mental integrity, cognitive autonomy should be strengthened through AI systems that enhance human agency, support well-being and empower informed decision-making. AI systems can also threaten cognitive autonomy, especially when they seek to manipulate emotions and beliefs, simulate friendship, or exploit psychological vulnerabilities. Developers and deployers must not design or operate AI systems that compromise cognitive autonomy or maximise engagement at the expense of people's well-being and freedom of thought.

Where neurotechnologies are involved, these principles should be complemented by the safeguards commonly described as neurorights.

10 Resilience

Technology should make society more resilient, never more brittle.

Resilience consists of protecting systems against failures and attacks while preserving the capacity of individuals, institutions and societies to continue functioning, adapt and thrive in an environment of accelerated technological change.

Resilience transforms innovation into sustainable progress and ensures that its benefits can endure through future crises and transformations.

11 Institutional, Public and Societal Capacities

Public capacity requires equipping institutions with sufficient knowledge to understand, assess, procure, oversee and, where appropriate, develop AI systems. Such capabilities represent an investment in democratic autonomy because they enable societies to decide how technology should be used.

12 Digital Infrastructures Serving the Public Interest

Public digital infrastructures form the structural foundation of inclusive, resilient AI that serves the common good. Investing in such infrastructures means investing in the collective capacity to innovate, compete and make autonomous decisions within the knowledge economy.

Public procurement should serve the same end, rewarding openness, interoperability, security and sustainability, and creating demand for trustworthy alternatives in the critical layers of the digital stack.

Public action does not replace private innovation; it creates the conditions under which it can thrive in an open and competitive environment.

Open and interoperable AI ecosystems, prioritising open-weight models, tools and standards, play a vital role in fostering innovation, scientific progress, transparency, competition and democratic access to technology. Public policies should support their development and adoption while ensuring appropriate safeguards for safety, security and fundamental rights.

Beyond physical infrastructure, the AI economy needs common foundations that are still largely to be built: interoperable standards; trustworthy identity for people, organisations and the AI agents acting on their behalf; provenance and trust signals for information; mechanisms that recognise and remunerate creators and other contributors; secure and resilient systems; and fair access to compute, data and models. Public institutions should help build these foundations together with researchers, civil society, workers, creators and enterprises of every size, so that they remain open to all.

Control over data should remain with those who produce or hold it. Wherever possible, AI systems should access data on its holders' terms, for a declared purpose and with a record of use, rather than extracting it in bulk. Distinguishing the custody of data from the ownership of models opens markets to many developers, protects privacy and creative work, and supports fair competition.

Because no one yet has all the answers, the governance of these foundations should be adaptive: monitoring effects, testing new approaches openly, learning from practice and sharing results across countries and communities.

SECTION III

INCLUSIVE, INNOVATIVE AND SUSTAINABLE ARTIFICIAL INTELLIGENCE

13 Legitimate Purpose

The use of AI must address a legitimate, specific and clearly identifiable need, provide demonstrable social benefit, and effectively contribute to human well-being, the protection of rights, or the achievement of public-interest objectives.

The requirement of a legitimate purpose not only guides the beneficial use of AI but also establishes clear limits on its deployment. Uses of AI that are inherently incompatible with human dignity, fundamental rights, and international peace and security cannot be considered legitimate.

Specifically, lethal autonomous weapons are morally repugnant, and no machine should ever choose to take a human life. There must be meaningful and accountable human control over all weapons systems, and a ban on lethal autonomous weapons.

14 Technological Proportionality

The use of AI systems must be proportionate to the needs they are intended to address, the risks they generate, and the resources they require.

Proportionality should be assessed not only in terms of expected benefits but also in relation to its effects on rights, autonomy, environmental sustainability, resource consumption, and the distribution of costs and benefits.

15 A Broad Concept of Progress

AI should contribute to a broad conception of human progress, understood not only as economic growth or technological advancement, but also as the expansion of human capabilities, the strengthening of rights, the improvement of well-being, social cohesion, environmental sustainability and democratic deepening.

The digital transformation should foster an inclusive and balanced distribution of the value created among all participants in the ecosystem of data, models, infrastructure, labour, services and territories. Collective bargaining, fair taxation and public stakes in AI infrastructure are key mechanisms for achieving this. Value should flow to those who contribute to it, including creators, researchers, workers, communities and enterprises whose knowledge and work AI systems draw upon, and not only to whoever controls the final point of distribution.

AI is far more than a race to build the best model; it is a race to build the society that is best able to harness and benefit from AI. Europe's competitiveness in the age of AI lies in its millions of small and medium-sized enterprises, start-ups, creators and problem-solvers, and in open, contestable markets in which all of them can build with AI. Clear, predictable and proportionate rules, accompanied by practical guidance, regulatory sandboxes and simple paths to compliance, give enterprises of every size the confidence to experiment, invest and grow.

16 Education and Training

Education and lifelong learning are fundamental pillars in the age of AI. Social integration demands empowering people. Education involves not only teaching individuals to use new tools but also strengthening the capabilities to think critically, distinguish information from manipulation, form independent judgments, and participate in social and democratic life in an informed manner.

Success in the AI era will depend less on possessing the most advanced technology than on integrating that technology effectively into society.

17 Enhancement of Work

Technological transformation should primarily aim to complement the capabilities of workers and professionals, improve job quality and foster more collaborative and effective organisations. Productivity gains from AI should be shared with workers through improved pay and working conditions, including opportunities to reduce working time without loss of pay.

Workers should be supported and protected throughout the digital transformation, with access to training, social dialogue, fair working conditions, measures to prevent avoidable job displacement and support redeployment and income security, and safeguards against disproportionate monitoring or adverse impacts arising from the use of AI. These protections should extend to workers throughout AI supply chains, including subcontracted workers, data annotators and content moderators.

This process is not determined by technology but by the organisational decisions that accompany it: who participates in its design and deployment, the objectives pursued, and how its benefits are distributed. These decisions will determine whether automation enhances professional capabilities.

Workers and their representatives should have effective rights to participate in decisions concerning the use of algorithmic systems at work, including through social dialogue, collective bargaining, and information and consultation processes.

18 Social, Linguistic and Cultural Diversity

AI should contribute to preserving and strengthening the linguistic, cultural and territorial diversity of societies. AI systems should appropriately recognise and represent different languages, cultures, knowledge traditions and forms of expression, broadening participation and preventing new forms of exclusion.

19 Intergenerational, Material and Territorial Sustainability

AI should contribute to a broad and enduring vision of sustainability. AI should generate prosperity, knowledge and well-being for present and future generations, promoting a more efficient and intelligent use of resources, increasing the resilience of our infrastructure, and supporting balanced territorial development while respecting natural resources and improving our ecological footprint.

Sustainability also requires that the physical footprint of AI, from the energy, water, land and materials consumed by data centres and hardware to the waste they produce, be measured, transparently disclosed, kept within environmental limits and be subject to ongoing community and environmental assessment.

Sustainability also has a financial dimension. As AI infrastructure is increasingly financed through a combination of debt, private credit and other financial vehicles, it is essential to ensure transparency regarding how risks and responsibilities are distributed. Public investment and partnerships in AI infrastructure should therefore be accompanied by clear disclosure of who ultimately bears those risks.

SECTION IV

RESPONSIBLE, TRANSPARENT AND TRUSTWORTHY ARTIFICIAL INTELLIGENCE

20 Identity and Transparency

Transparency must be operational, enabling an understanding of the system's purpose, limitations, risks, relevant data, operating criteria, degree of human intervention and available channels for submitting complaints and seeking redress.

As AI systems become increasingly capable of acting autonomously, trustworthy digital societies require interoperable identity infrastructures that make it possible to identify who is acting, on whose behalf, and under what authority. AI agents should be uniquely identifiable and their actions attributable, while responsibility for those actions must always remain with accountable natural or legal persons.

Democratic societies should promote open, interoperable and privacy-preserving identity standards that support trust, security and accountability across digital ecosystems.

21. Trustworthiness and reliability

Trust makes the widespread adoption of AI possible, as it encourages individuals, organisations and society to embrace its use when it is evidenced as safe, fair and understandable. Trust is warranted only when it follows demonstrated reliability, so the aim must be AI that is trustworthy. Where AI is trustworthy, people and organisations can build upon it with confidence, and its benefits can spread across the whole economy.

Reliability requires acknowledging the limitations of AI and ensuring an effective capacity to correct, remediate and withdraw systems whenever necessary.

Reliability must be assessed regarding errors, uncertainties, lack of context and system limitations, from design through to decommissioning.

22. Traceability, Auditability and Redress

AI systems must be traceable with respect to their data, models, outputs, responsible actors, updates and incidents.

Traceability must enable errors to be detected and corrected, harms to be identified and remedied, relevant decisions to be reconstructed, compliance with requirements to be verified, accountability to be established and corrective measures to be implemented.

Systems must include clear mechanisms for suspension, modification, contestation or withdrawal whenever their outcomes compromise safety or rights.

Responsibility for harm caused by AI systems must rest with those who develop and deploy them: where high-risk systems cause harm, the burden of proof must shift to developers and deployers, affected people must be able to seek redress collectively, and those who report risks or wrongdoing from within must be protected.

23. Assessment Throughout the Entire Lifecycle

The assessment of AI systems should accompany them throughout their lifecycle, from design and development to deployment and use, ensuring that they remain safe, reliable, useful and consistent with the objectives and values that justified their adoption.

Such assessment benefits from cooperation among system developers, independent evaluators, public authorities, workers, users and affected communities.

Frontier and agentic AI systems must undergo independent evaluation before deployment, with evaluators given access to models, training-data documentation and incident records. Public authorities, not only operators, must have the power to suspend or withdraw any system whose behaviour escapes effective human control.

24. Meaningful Human Control

Human oversight must be real and effective, supported by the knowledge, context, information, time, authority and genuine capacity to understand, correct, reject or halt AI-generated outcomes. The growing autonomy of agentic AI increases the need to ensure robust mechanisms of human oversight. These requirements apply equally to public authorities, whose use of AI, particularly in areas such as security, welfare and justice, calls for the highest standards of transparency, contestability and human control.

Where AI agents act on behalf of people or organisations, every action they perform must remain attributable to an accountable principal through interoperable and privacy-preserving infrastructures for identity and mandates.

25. Justice and accountability

AI should contribute to making justice more accessible, effective and transparent, strengthening accountability in all decisions that affect people. Its use in this field must be grounded in respect for the rule of law, due process and verifiable facts.

AI systems may support technical and administrative functions, but judicial decisions and decisions that significantly affect people's rights must remain subject to meaningful human oversight and accountability.

Closing Statement

This Manifesto expresses a shared belief: artificial intelligence must serve the interests of all humanity. We support clear limits, effective accountability and the fair distribution of AI's benefits, while protecting human rights, autonomy and dignity. Our aim is not only to govern AI responsibly, but to create the conditions under which people, communities, researchers, creators and enterprises of every size choose if, how, and where they build with AI in ways that strengthen democracy, expand opportunity and contribute to shared prosperity. Where AI is trustworthy and its foundations remain open and accessible, its benefits can spread across the whole economy and society.

