Task Force 4
Digital Transformation

Policy brief

HUMAN-CENTRIC AI: FROM PRINCIPLES TO ACTIONABLE AND SHARED POLICIES

SEPTEMBER 2021

Carlo Casalone Pontifical Academy for Life
Luciano Floridi University of Oxford
Laura Palazzani Pontifical Academy for Life
Renzo Pegoraro Pontifical Academy for Life
Francesca Rossi IBM Research
Roberto Villa Fondazione IBM Italia

T20 NATIONAL COORDINATOR AND CHAIR
ISPI

T20 CO-CHAIR
IAI

T20 SUMMIT CO-CHAIR
Università Bocconi MILANO

THINK20 ITALY 2021

G20 ITALIA 2021

ABSTRACT

AI is increasingly being used to support and improve human decision-making. This technology holds the promise of delivering valuable insights and knowledge across a multitude of applications. However, the broad adoption and positive impact of AI systems will rely heavily on the ability to trust the whole AI ecosystem, insofar as it is capable of promoting the autonomy of human beings while recognising and respecting their fundamental vulnerability.

To achieve this, we recommend an ethics-by-design methodology for adoption by G20 Governments, which can drive the design, development, and deployment of trustworthy AI/digital ecosystems in each State. This methodology should be harmonised and supported by a multi-national approach inspired by the recognition of trustworthy AI as a common good.

CHALLENGE

As Artificial Intelligence (AI) systems enable significant transformations in social practices and lifestyles, our mental and interpersonal habits are undergoing profound changes, putting human centrality, dignity, and agency in jeopardy. Also, the rapid deployment of AI and the different socio-cultural contexts can amplify global digital divides, with a potentially negative impact on fairness, social justice, and inclusivity. Another challenging dimension regards the impact on our planet: advanced technology, like AI, provides new opportunities to tackle the urgency of the sustainability challenge. However, this potential has not yet been fully leveraged, putting billions of people and “our common home” at risk.

While previous societal transformations, such as the one generated by the industrial revolution, have also been addressed by a widespread education system, the existing education system now struggles to cope with these challenges. This is particularly true in vertical STEM education, which provides only technical/digital skills, without humanistic insight.

In recent years, Governments have published national AI strategies and regulations, enterprises have defined their principles for a beneficial AI – and people, media, and all other stakeholders have engaged in AI ethics discussions and initiatives (see, for example, the recent AI Regulation proposal by the European Commission). At this point, awareness is growing that a human-centric multi-stakeholder approach for AI is needed to assure a sustainable and inclusive future, even if there is some divergence about what human-centric AI means in practice, and what policies should be put in place to achieve it.

Our proposal supports the G20 Governments in addressing this challenge, by providing a methodology that they can effectively use to move from AI strategic plans to concrete and shared activities. These include defining shared terms, setting goals, identifying maturity attributes, performing measurements, setting up ethics committees, agreeing on AI regulatory approaches, and employing assessment criteria. The overall goal is to increase the trustworthiness and beneficial impact of the whole AI ecosystem in their respective countries and globally.

PROPOSAL

In order to define shared actionable policies around AI, we first propose to establish a *Foundational Framework* to identify the relevant AI-related pillars, goals and metrics, and then to employ an *Operational Framework* for strategic planning and effective execution.

The Foundational Framework relies on the concept of a human-centric AI ecosystem, which includes dimensions such as meaningful human control, transparency, explainability, fairness, justice, inclusiveness, sustainability, and education. To define these dimensions properly, it is necessary to combine pragmatic and technological considerations with theoretical, philosophical, and scientific ones (see Unesco-Comest 2019, p. 5.) We also suggest that the conceptual framework of fundamental human rights can provide a shared framework for both national and international initiatives (see EC High-Level Experts Group on AI 2019; EGE 2019).

Meaningful human control: in a human-centric approach to AI, it is necessary to understand, define, and regulate the synergy between human beings and machines (which should be interpreted as a complement and support rather than a replacement) seeking interaction modalities that allow humans to maintain significant and meaningful control in terms of intervention, supervision, and responsibility. This also implies possible legal solutions that rule out the possibility of machines being recognised as individuals or moral agents, or being attributed an electronic personality (see European Parliament, 2017 Recommendations). Instead, AI systems should be categorised as things (that is, able to perform without awareness).

Transparency and explainability: given the complexity of the algorithms and the large amount of data processed by AI, it is not easy to ensure that humans are in control, especially without an effective form of explainability that clarifies how the system reaches its decisions and hence supports transparency and human oversight. The issue at stake is to bridge the gap between the complexity of the process, which usually needs to handle a large amount of data, and the human ability to remain in control.

Fairness and justice: it is important to recognise situations of inequality and differences between opportunities and resource infrastructures at international and the national level. AI should be used to increase fairness and reduce/eliminate discrimination and lack of opportunities. Governments should define metrics to assess the impact of AI on these aspects and aim to improve on them, including through the use of AI itself.

Inclusiveness: all stakeholders should be involved in planning, developing, deploying, and regulating AI. Particular attention should be paid to the Global South, people living below the poverty threshold, marginalised communities, persons with disabilities, and religious and ethnic minorities.

Sustainability: AI systems should benefit all human beings, including future generations. Their sustainability should be ensured, both on a social and an environmental level. Particular attention should be paid to the social and environmental impact of AI. AI technology should be based on human responsibility to ensure healthy conditions for life on our planet, preserving a suitable environment for future generations.

Education: the study of humanities must be incorporated into scientific and technical disciplines, to enable people to understand the responsibilities involved in the development, deployment and use of AI.

Education should be revised, introducing ethics into curricula for engineers, IT experts, computer technicians, computer scientists, and data scientists, with particular reference to data and technology ethics. This would ensure an awareness and understanding of ethics from the earliest stages of technological design, in order to anticipate ethical issues before deployment. AI designers and developers both need help developing a critical awareness of AI and understanding the capabilities, limitations, and risks of technologies to guarantee their ethical design, deployment, and use.

Education should also include programmes of lifelong learning aimed at people already in the workforce, to avoid skills polarisation and deskilling, and ensure a re- and up-skilling trend able to promote digital capabilities and ethics awareness to address the development of new AI technologies.

Education should provide public information to ensure basic AI literacy and also cover ethical aspects involved in applying the technology, promoting active participation in social discussion to support and ensure inclusiveness.

FROM THE FOUNDATIONAL TO THE OPERATIONAL FRAMEWORK

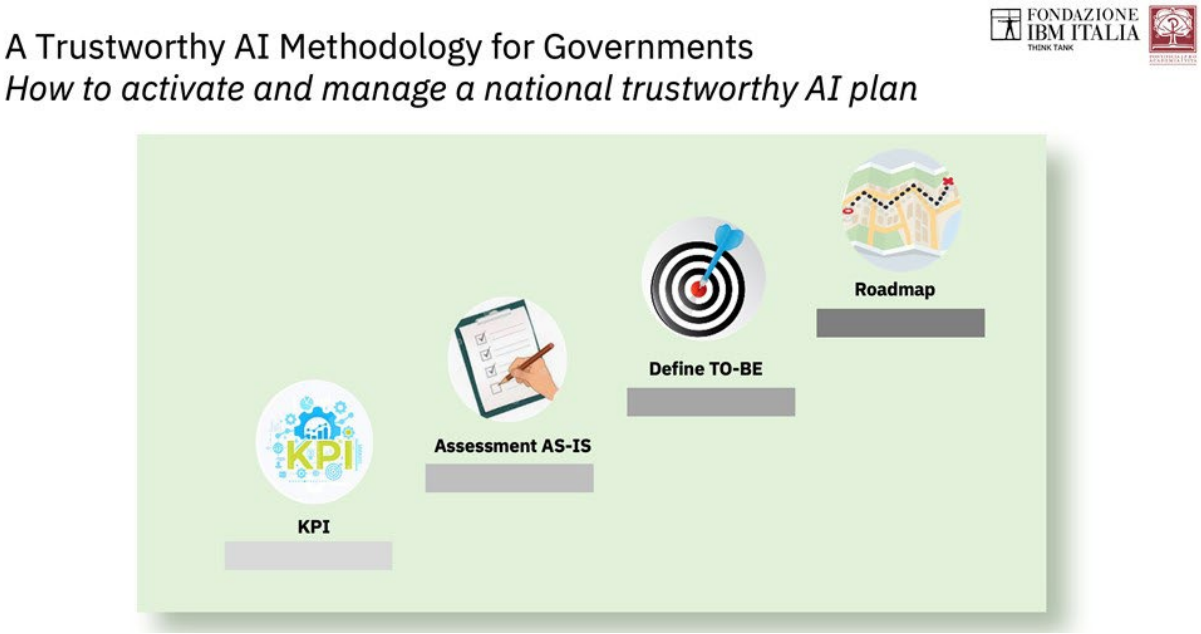
The Foundational Framework provides the main guidelines for the design and the development of trustworthy AI/digital ecosystems. However, differences between values and priorities in the various States may lead to significant differences in framework implementation. To support a consistent rollout across States, we recommend several coordinating activities.

The first one is to develop a standard glossary of foundational human-centric AI/digital ecosystem terms and pillars, that needs to be collectively defined and adopted by all States, to ensure a shared understanding of the main concepts, pillars, and values.

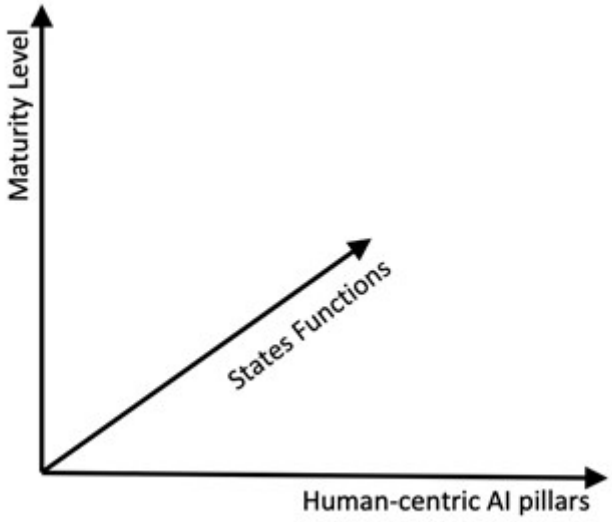
The Operational Framework provides G20 Governments with the methodology to drive the design and deployment of human-centric AI/digital ecosystems, which is dependent on implementing the Foundational Framework described above. The effectiveness of the Opera-

tional Framework will depend on a continuous assessment of how human-centric AI pillars are embedded and adopted in each Government system.

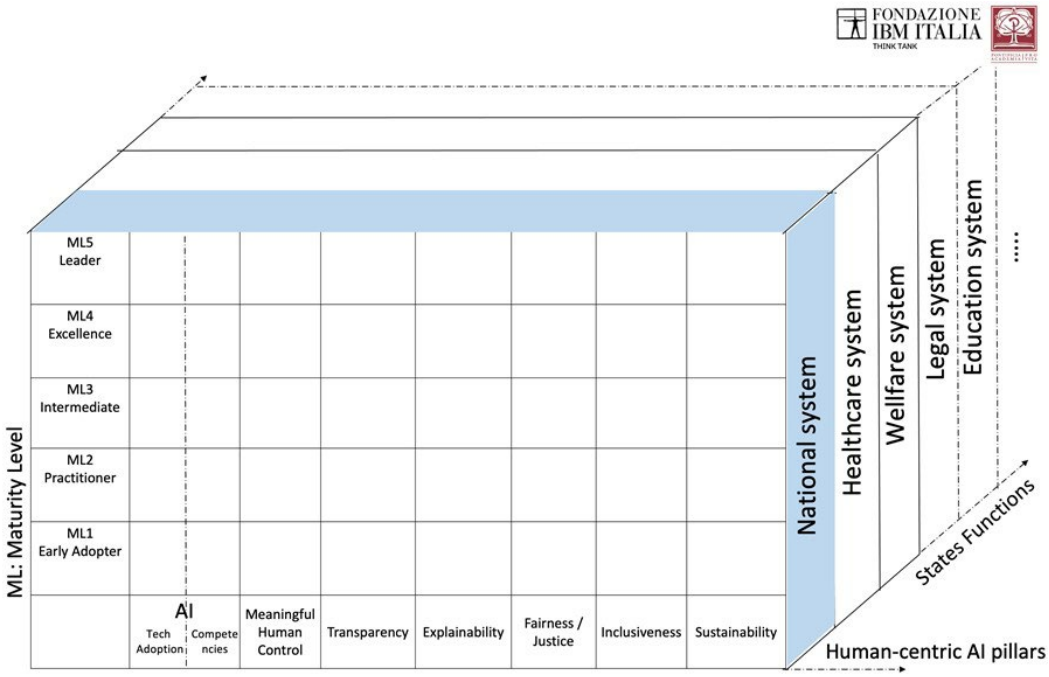
We propose that the Operational Framework should include a *Trustworthy AI Methodology* for Governments that provides a multi-stakeholder, multi-country, and multi-cultural approach to create a system of trust in technology and its uses, based on a multi-dimensional, risk-based AI governance framework that includes education, training, toolkits, methodologies and their adoption, governance models, ethics committees and best practices. This methodology should be used by individual countries and multi-country coalitions (such as the G20).



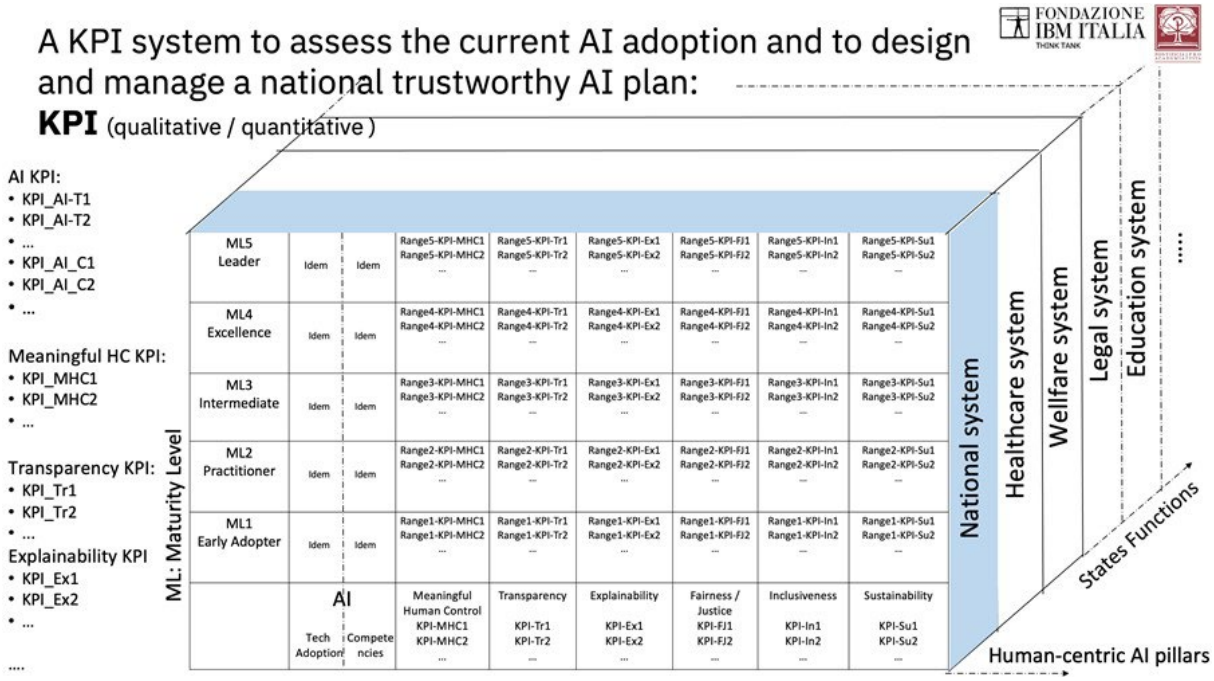
A three-dimensional matrix is used to model AI adoption by a State. The X-axis represents the human-centric AI pillars, the Y-axis represents the maturity scale, and the Z-axis represents State functions (i.e. the healthcare system, the education system, etc.).



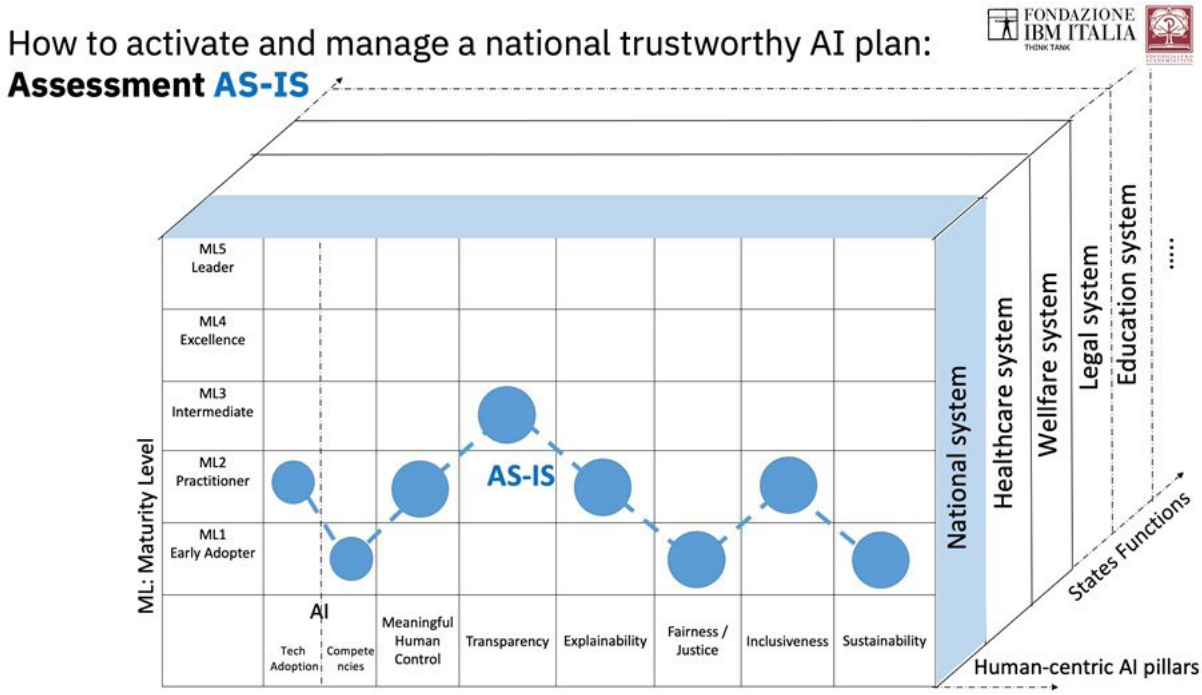
The following is an instance of the three-dimensional matrix:



Application of the methodology requires the selection of Key Performance Indicators (KPI) to assess the current status of the AI adoption and subsequent tracking of action plan progress so that improvements can be made. A supranational organisation like G20 should determine KPIs to ensure that they are consistent across countries. KPIs can be quantitative and qualitative, and are specific to each human-centric AI pillar. An example of a KPI for the Transparency pillar is “public sharing of the data used for the training of algorithms”. For instance, if a Government has deployed many algorithms and the data used for training is not shared in most cases, this is symptomatic of a low Transparency maturity level.

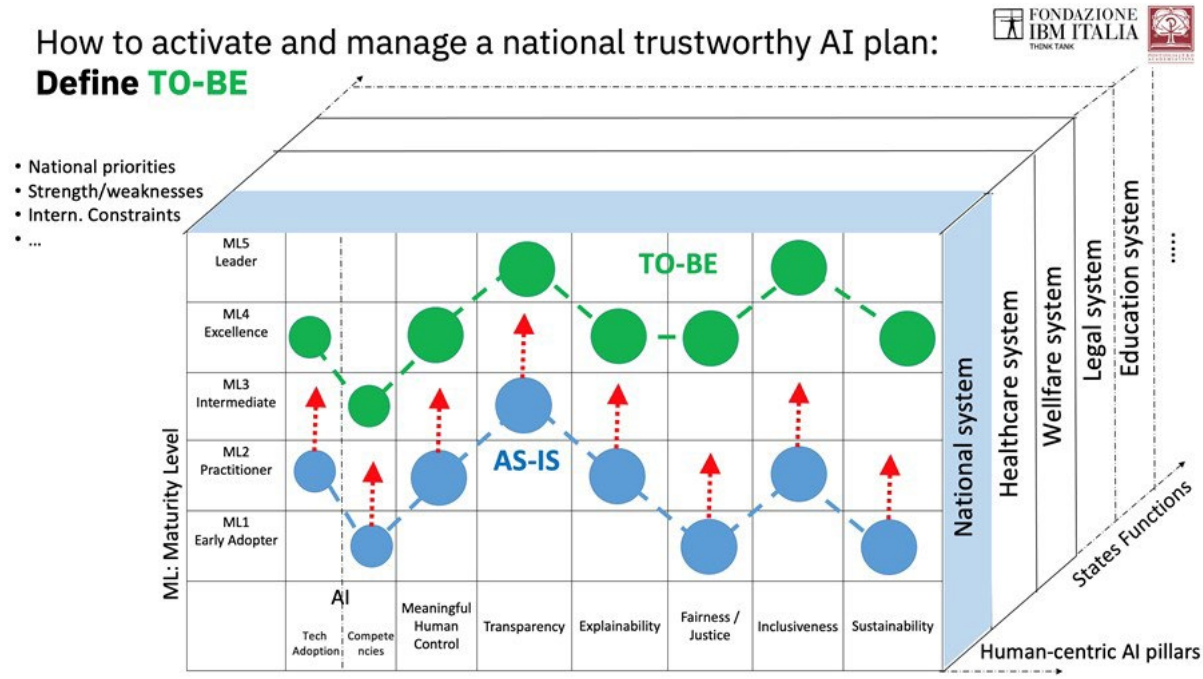


Governments use the KPIs to assess current AI adoption status (AS-IS):



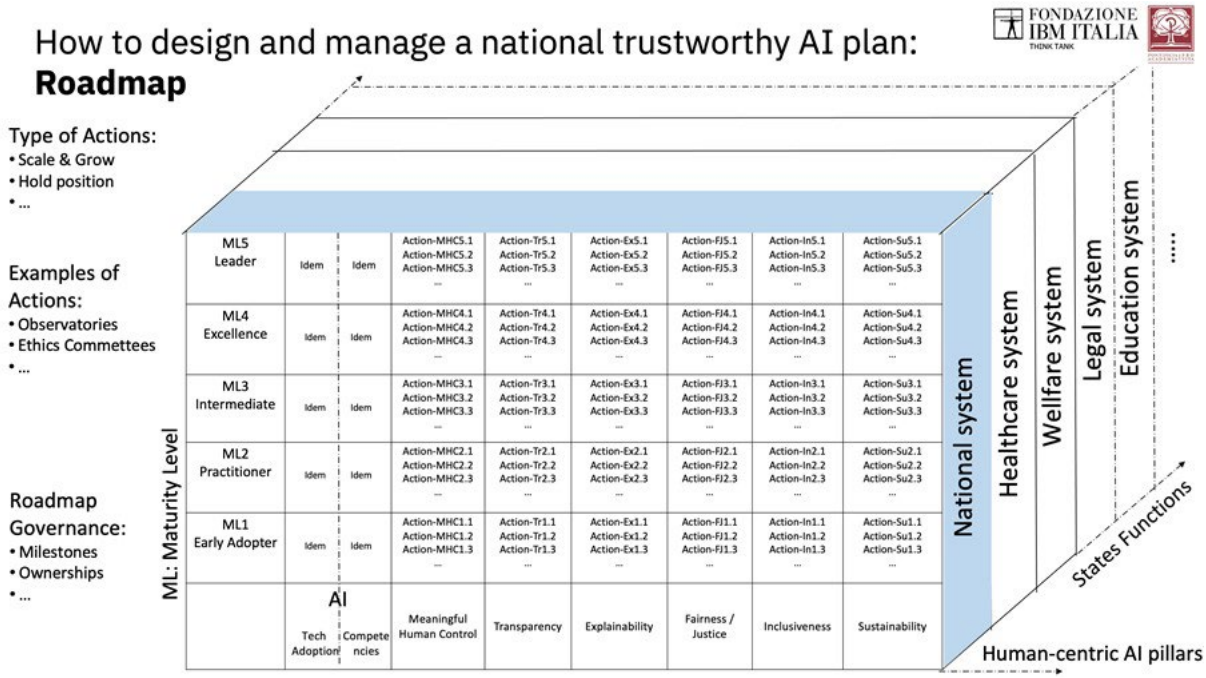
The AS-IS matrix provides a snapshot of: (1) current status at national level and at the level of functions/systems such as Healthcare, Welfare, Legal, Education and so on (State Functions axis); (2) AI adoption and competency maturity level (ML axis); and (3) human-centric AI pillar adoption (x-axis). The AS-IS is the outcome of the assessment carried out using the KPIs, applied to current AI adoption status.

The next step is to define the target AI adoption status (TO-BE):



When the current status (AS-IS) is defined, each State can define its target status (TO-BE) based on its own national priorities, strengths and weakness, and strategic goals. The target status (TO-BE) can be defined at national or systems level.

To move from AS-IS to TO-BE status, each State must define a ROADMAP with the actions needed to make the expected improvement, including milestones and ownerships. The KPIs will be regularly monitored to track progress over time.



Each element of the human-centric pillars (i.e., transparency, fairness, etc.) can move to a higher level of maturity if a set of activities aiming at that goal is in place. The activities are specific to each element and its maturity level.

When applied to a State, the methodology should work across the nation's multiple systems (Healthcare, Legal, Wellness, Education, etc.) and address critical problems in that system while providing the capability to assess and increase their Maturity Level.

The kind of actions that we recommend including in the Operational Framework include:

- setting up specific interdisciplinary and independent AI *Ethics Committees* and *Observatories*, at national and international level;
- implementing *ethics monitoring* throughout the algorithm's life cycle;
- learning from the *best practices* of other States/Governments in the coalition, to leverage the valuable work already done by similar nations.

CONCLUSION

THIS IS A SUMMARY OF RECOMMENDATIONS FOR G20 GOVERNMENTS AND THE G20 AS A MULTI-GOVERNMENT ORGANISATION

Recommendations to individual Governments:

1. define human-centric AI in terms of meaningful human control, transparency, explainability, fairness, justice, inclusiveness, sustainability, and education. Combine technological and philosophical considerations. Adopt a fundamental human rights framework.
2. endorse the OECD AI principles.
3. interpret AI systems as a support to human decision-making, not a replacement. Do not recognise machines as moral agents and do not give them an electronic personality or identity.
4. require explainability and transparency in AI systems.
5. define metrics to assess the impact of AI on fairness and social justice, and strategic plans to improve such metrics.
6. apply a multi-stakeholder approach to all decisions regarding AI.
7. measure the impact of AI on the environment. Consider the well-being of both current and future generations when deciding on AI-related initiatives, incentives, funding, and policies.
8. include data and technology ethics in science curricula. Expand lifelong learning initiatives. Create AI literacy activities for citizens.
9. set up an independent and multi-disciplinary AI ethics committee, in each Government and at G20 level.
10. when regulating AI, impose conditions on the uses of AI (and not AI per se), and adopt a non-territoriality approach whereby the rules of a specific country apply to whoever deploys AI in that country.

Recommendation to G20 organisation: set up a G20 agency (G20-AI) that defines shared initiatives around AI. Recommendations to G20-AI:

1. define a standard glossary including all aspects of human-centric AI.
2. define shared KPIs to assess current AI adoption along three axes (human-centric AI pillars, State functions, maturity level).
3. set up an independent and multi-disciplinary AI ethics committee, including representatives of all 20 country-level AI ethics committees.
4. identify ways for more AI-mature countries to support and accelerate the journey taken by less AI-mature countries towards human-centric AI.
5. Define and share milestones and timelines for adopting and implementing the operational approach for all Governments.

REFERENCES

EU document "Ethics Guidelines for a Trustworthy AI" https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

EU document "Policy and Investments Ethics Guidelines for a Trustworthy AI" https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

European Commission, *White paper on AI. A European Approach to Excellence and Trust*, 19 February 2020

European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence", 21 April 2021

European Group in Ethics in Science and New Technologies, *Statement on Artificial Intelligence, Robotics and 'Autonomous Systems'*, 2018

European Parliament, *Resolutions with Recommendations to the Commission on Civil Law Rules on Robotics*, 6 February 2017

Floridi L. et al. "A.I.4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *Minds and Machines*, 2018

High-Level Experts Group on AI, *Ethics Guidelines for Trustworthy Artificial Intelligence*, 8 April 2019

Italian Committee for Bioethics, together with CNBBSV, *AI and Medicine: Ethical Aspects*, 2020

Nuffield Council on Bioethics, *Artificial Intelligence in Healthcare and Research*, 2020

Rome Call for AI Ethics, Città del Vaticano, 28 febbraio 2020.

OECD, *Legal Instruments Artificial Intelligence*, 2020

The Future Society, "Areas for future action in the responsible A.I. ecosystem", December 2020

Paglia V. and R. Pegoraro, "The Good Algorithm? Artificial Intelligence: Ethics, Law, Health", *Pontifical Academy for Life*, 2020

E. Sinibaldi et al. "Contributions from the Catholic Church to ethical reflections in the digital era", *Nature Machine Intelligence*, 2020

UNESCO, COMEST, *Robotics Ethics*, 2018

UNESCO, COMEST, *Preliminary Study on the Ethics of Artificial Intelligence*, 26 February 2019

Educational material for designers and developers:

Everyday Ethics for AI: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

External articles:

Harvard Business Review article, 2020: https://hbr.org/2020/11/how-ibm-is-working-toward-a-fairer-ai

External partnerships:

Partnership on AI: https://www.partnershiponai.org/

Global Partnership on AI: https://gpai.ai/

IEEE AI ethics initiative: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Word Economic Forum Global Future Council on AI for Humanity: https://www.weforum.org/communities/gfc-on-artificial-intelligence-for-humanity

Notre-Dame Tech Ethics Lab: https://techethicslab.nd.edu/

Global studies:

IBM IBV study on "Advancing AI ethics beyond compliance": https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics

IBM approach to AI Ethics:

IBM AI Ethics web site: https://www.ibm.com/artificial-intelligence/ethics

Trusted AI for business https://www.ibm.com/watson/ai-ethics/

Open-source toolkits:

AI fairness 360: https://aif360.mybluemix.net/

AI explainability 360: https://aix360.mybluemix.net/

AI factsheet 360: http://aifs360.mybluemix.net/

Public policies:

IBM Policy Lab: https://www.ibm.com/policy/

AI precision regulation: https://www.ibm.com/blogs/policy/ai-precision-regulation/

Facial recognition: https://www.ibm.com/blogs/policy/facial-recognition/

Response to COVID-19: https://www.ibm.com/thought-leadership/covid19/

ABOUT THE AUTHORS

Carlo Casalone Pontifical Academy for Life, Rome (Italy)

Carlo Casalone, M.D., S.T.D. is a Visiting professor in Moral Theology and Bioethics at the Pontifical Gregorian University of Rome (Italy) and coordinator of the Scientific section of the Pontifical Academy for Life. He has been a counselor of the Pontifical Council for Healthcare workers and a member of the Ethics Committee for Clinical Research of the National Cancer Institute, Milan, and has been teaching as an Associate professor at the Pontifical Theological Institute of Southern Italy.

Luciano Floridi University of Oxford (UK)

Professor of Philosophy and Ethics of Information at the University of Oxford and Professor of Sociology of Culture and Communication at the University of Bologna, where he is Director of the Center for Digital Ethics. He is a world famous expert in digital ethics, artificial intelligence ethics, philosophy of technology and philosophy of information.

Laura Palazzani Pontifical Academy for Life, Rome (Italy)

Full Professor of Philosophy of law at Lumsa University, Rome. Corresponding Member of the Pontifical Academy for Life. Deputy Vice-chair of the Italian National Bioethics Committee at the Presidency of the Council of Ministers. Italy's official delegate to the Committee on Bioethics DH-BIO, Council of Europe. Member of the UNESCO International Bioethics Committee. 2010-2021 member of the European Group on Ethics in Science and New Technologies, European Commission

Renzo Pegoraro Pontifical Academy for Life, Rome (Italy)

Chancellor of the Pontifical Academy for Life in Rome. Professor of Bioethics at the Faculty of Theology of Northern Italy. Member of the European Society of Philosophy of Medicine and Health Care (ESPMH).Professor of Nursing Ethics at Children Hospital "Bambino Gesù", Rome. Member of the Board of the European Association of Centres of Medical Ethics (EACME). Member of the International Association for Education in Ethics (IAEE).

Francesca Rossi IBM Research (Italy)

IBM Fellow and AI Ethics Global Leader. Co-chair of the IBM AI ethics board. Board member of the Partnership on AI. Member of the Steering Committee of the Global Partnership on AI. AAAI (Association for the Advancement of AI) and EurAI (European Association of AI) Fellow. President-elect of AAA

Roberto Villa Fondazione IBM Italia

IBM Global University Programs Leader for Europe, IBM Research. Member of the Technical and Scientific Committee of MADE, the Competence Center Industry 4.0 led by Polytechnic University of Milan Member of the steering group of StudENT for Africa, led by University of Bologna, to foster entrepreneurship development in Africa. Lecturer on the human centric approach for technology, research and business at European universities.