Grant Agreement Number: 957204 (H2020-ICT-38-2020)
Project Acronym: MAS4AI
Project Start Date: 1st October 2020
Project Full Title: Multi-Agent Systems for Pervasive Artificial Intelligence for assisting Humans in Modular Production

D6.6 - Ethics framework

Dissemination level:	PU(Public)
Date:	2021-30-09
Deliverable leader:	US
Contributors:	G. Bar
Reviewers:	DFKI
Type:	R
WP / Task responsible:	WP6 / Task 6.6: Ethics framework
Keywords:	Legal AI, Trustworthy AI, Ethical AI, Explainable AI, Transparency, Human oversight, Autonomy, Cybersecurity

Executive Summary

Artificial Intelligence (AI) is having a huge impact in many areas of societal life, including the industry. Many different organisations around the world have launched a wide range of initiatives to establish ethical principles for the adoption of ethical AI. Due to the ethical, legal and privacy risks posed by the wide application of AI in smart factories, it is crucial that the MAS4AI pilots are consistent with widely adopted rules or existing laws. This document contains an initial analysis of ethical framework.

Document History			
Version	**Date**	**Contributors**	**Description**
0.1	17.09.2021	G. Bar	Outline
1.0	30.09.2021	A. Girenko	Review

Table of Contents

Table of Figures

1 Main ethical challenges

1.1 Biases and data quality

The main challenge for AI/ML algorithms is to have good quality data sets to eliminate bias in that data. High-quality training, validation, and test datasets require the implementation of appropriate data management practices. Data sets should be adequate, representative, error-free and complete from the point of view of the purpose of the system. They should also have appropriate statistical characteristics, including in relation to the persons or groups of persons for whom the high-risk AI system is to be used.

1.2 Unequal treatment and discrimination

AI systems have the potential to create and strengthen unequal treatment, including bias in underlying data sets, and thus create various forms of discrimination, including indirect discrimination, in particular with regard to groups of people with similar characteristics. AI technologies should be designed to respect cultural and linguistic diversity and to help meet basic human needs. Any use that might discriminate on the basis of age, sex, sexual orientation, beliefs or other characteristics of sensitive data nature, would risk prejudicing physical or mental autonomy, or would lead to unjustified surveillance or manipulation, must be avoided.

Particular attention should be paid to the possibility of discrimination in employment relations. In the recruitment process and employee evaluation and promotion process, AI systems can perpetuate historical patterns of discrimination against, for example, women, certain age groups, people with disabilities, or people of certain racial or ethnic origin or sexual orientation.

1.3 Privacy threats

The use of huge data sets and the speed of their processing pose a challenge to privacy protection. There is no doubt that the more data for AI training, the better the results, but each time it is necessary to answer the question about the origin of this data, the purpose of its use and the legal basis for it.

In particular, remote recognition technologies such as the recognition of biometric identifiers (e.g. facial recognition) pose particular privacy risks. Their use should always be disclosed, proportionate, targeted and limited to a specific purpose, limited in time and carried out in accordance with EU law, including GDPR, with due respect for human dignity and autonomy and fundamental rights.

It should be emphasised that AI systems used to monitor employee performance and behaviour can significantly affect their right to data protection and privacy.

1.4 Power imbalance

There is a clear asymmetry between the actors using artificial intelligence technologies and the actors interacting with and affected by these technologies.

In this context, it is important to ensure the explainability of AI systems and the possibility of appealing against decisions made by them towards individuals.

1.5 Opacity/transparency imbalance

Still one of the bigger challenges for AI is transparency regarding interaction with AI, how AI system works, what capabilities it has, how information is filtered and presented, what is accuracy of results and the system limitations.

Users' trust is essential to the development and deployment of technologies that may have inherent risks if they rely on opaque algorithms and biased datasets. AI users should have the right to be adequately informed in an understandable, timely, standardised, accurate and accessible manner about the existence of algorithmic systems, the reasoning they use, their possible effects and consequences for users, how to reach decision-making people and how to control, effectively challenge and correct system decisions. Ensuring the transparency of AI systems also applies to such issues as: who is responsible for AI and owns the results of AI work.

1.6 Environmental impact

The computer centres that run infrastructure for data collection and learning AI are very power-hungry. Due to their significant environmental impact, the carbon footprint of these technologies should be monitored throughout their life cycle, including the consumption of essential raw materials, energy and greenhouse gas emissions. The number of AI solutions aimed at ensuring environmental protection should also be increased.

1.7 Threat to human labour

One of the key problems with AI for industry workers is ensuring that these technologies serve, not replace, people. The ultimate goal of implementing AI solutions should be to increase the well-being of everyone, not just maximise the company's profits.

It is also important to provide employees with appropriate training and teach them to work with intelligent machines in order to manage digital transformation with dignity and respect for humans.

1.8 Autonomy and control

AI makes increasingly important decisions. The problem is that AI often has to make split-second decisions, especially with regard to autonomous vehicles and production robots. A significant challenge is to find a balance between the autonomy of AI and necessary human supervision and

control. There is no doubt that increasingly autonomous robots should be able to perform their functions in complex environments in cooperation with human workers.

1.9 Vulnerabilities and cyber threats

Due to the significant impact that AI systems have on the environment in which they are used, care should be taken to ensure that they are precise and technically reliable. It is necessary to implement solutions to prevent security breaches, leakage and "poisoning" of data, cyber-attacks and the misuse of personal data. From the design stage, AI systems should be developed in a safe, traceable, technically reliable, ethical, legal way and should be subject to independent scrutiny and supervision.

1.10 Responsibility and accountability

All physical or virtual activities based on AI systems, devices or processes in which these systems are used, may be a direct or indirect cause of harm, and at the same time they are almost always the result of the fact that someone constructed or implemented such a system or interfered in it. Therefore clear rules for assigning responsibility for the actions of the AI system and fair compensation procedures are necessary to build the trust in AI. Any user who has suffered damage as a result of the operation of an AI should be able to successfully claim the compensation. Moreover, users should be sure that the potential damage caused by AI-powered systems is adequately insured and that there is a specific legal avenue for redress.

2 An overview of the key ethical principles for Artificial Intelligence

This report base on four key documents (guidelines, recommendations) regarding ethical AI:

1) The Asilomar AI Principles, developed under the auspices of the Future of Life Institute, in collaboration with attendees of the high-level Asilomar conference of January 2017 (hereafter **Asilomar AI Principles**);

2) The General Principles offered in the second version of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, published in December 2017 (hereafter **Ethically Aligned Design**);

3) European Commission High-Level Expert Group on AI - Ethics Guidelines for Trustworthy Artificial Intelligence, published in April 2019 (hereafter **HLEG Trustworthy AI**);

4) OECD Recommendation of the Council on Artificial Intelligence published in May 2019 (hereafter **OECD AI Recommendation**).

Below are presented the basic ethical principles formulated in the above-mentioned documents, which should be the basis for the development of the ethical framework for MAS4AI pilots.

2.1 Asilomar AI Principles

- **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
- **Failure Transparency**: If an AI system causes harm, it should be possible to ascertain why.
- **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
- **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
- **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation.
- **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

- **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyse and utilise that data.
- **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
- **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- **Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
- **Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
- **AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.

2.2 Ethically Aligned Design

- **Human Rights:** AI shall be created and operated to respect, promote, and protect internationally recognized human rights.
- **Well-being:** AI creators shall adopt increased human well-being as a primary success criterion for development.
- **Data Agency:** AI creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
- **Effectiveness:** AI creators and operators shall provide evidence of the effectiveness and fitness for purpose of AI.
- **Transparency:** The basis of a particular AI decision should always be discoverable.
- **Accountability:** AI shall be created and operated to provide an unambiguous rationale for all decisions made.
- **Awareness of Misuse:** AI creators shall guard against all potential misuses and risks of AI in operation.
- **Competence:** AI creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

2.3 HLEG Trustworthy AI (EU Approach to AI)

- **Human agency and oversight:** AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches

- **Technical Robustness and safety:** AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

- **Privacy and data governance:** besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.

- **Transparency:** the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations

- **Diversity, non-discrimination and fairness:** Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

- **Societal and environmental well-being:** AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.

- **Accountability:** Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.

2.4 OECD AI Recommendations

- **Inclusive growth, sustainable development and well-being:** Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes

for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

- **Human-centred values and fairness:** AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

- **Transparency and explainability:** AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

    - to foster a general understanding of AI systems,
    - to make stakeholders aware of their interactions with AI systems, including in the workplace,
    - to enable those affected by an AI system to understand the outcome, and,
    - to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

- **Robustness, security and safety:** AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

- **Accountability:** AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

3 Recommended ethical principles to be implemented in MAS4AI pilots

3.1 Well-being for environment and society

The principle of creating AI technology that is beneficial to society and environment is expressed in different ways across majority of ethical AI recommendations. One of the main pilot requirements to consider is to design and implement an AI system that is **sustainable and environmentally friendly.**

The social impact of AI, which may affect **people's physical and mental well-being**, should also be taken into account. Therefore, the effects of these systems must be carefully monitored.

The prominence of beneficence firmly underlines the central importance of promoting the well-being of people and the planet with AI[1].

3.2 Privacy

Pilots should prevent breaches of privacy and avoid misusing AI technology in any other way that may be harmful to personal data. AI systems must guarantee **privacy and data protection throughout an AI system's entire lifecycle.** This includes:

- the information initially provided by the user,
- the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations)[2].

Digital records of human behaviour, also in the shop floor, can enable AI systems to infer about the health, age, gender, religious or political views of data subjects. In accordance with the principle of minimisation in the GDPR, non-performance related data must be anonymised accordingly. Data may be also captured by sensors from both machines and operators. Sensors may record physiological parameters of operators (e.g., heart-rate, blood pressure, movement, etc.), the actions taken by operators (e.g. cameras or motion sensors) and **all of these will be personal data** (moreover- some of them will be sensitive data).

[1] L. Floridi, J. Cowls, *A Unified Framework of Five Principles for AI in Society*, Harvard Data Science Review, Issue 1.1, Summer 2019, p. 6.

[2] High-Level Expert Group on Artificial Intelligence (AI HLEG), Ethics Guidelines For Trustworthy AI, 8 April 2019, p. 16, https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai, p. 17.

It should be noted that any organization dealing with the data of natural persons should have rules regulating access to data. These rules should define who can access the data and under what circumstances.

3.3 Robustness, safety and security

AI system attacks can be against data (data poisoning), model (model leakage) or underlying infrastructure, both software and hardware. With regard to AI systems, cyber-attacks can be particularly dangerous as the data and behaviour of the system can be altered, leading to faulty decisions by the system that are harmful to the company or projct and/or users.

For AI systems to be secure, consideration should be given to possible unintended uses of the AI system (e.g. dual-use applications) and potential system abuse by malicious actors. The security of an AI system also means its **accuracy, reliability and reproducibility**[3].

3.4 Autonomy and Human oversight

One of the reasons AI is developed and implemented is to delegate some of our decision-making power to it. On the other hand, it is important to maintain a **balance between the decision-making power preserved for man and that which we give to machines.** The principle of autonomy and human oversight is one of the key ethical principles that runs through most of the documents on AI.

The autonomous systems cannot restrict the freedom of human beings to set their own standards and norms. In addition, it is people who should choose how and whether to delegate decisions to AI systems to achieve goals chosen by people. This means that both human autonomy should be promoted and the autonomy of machines should be limited and internally reversible if human autonomy needs to be protected or restored (e.g. the possibility for the operator to switch off the intelligent machine and regain full control of the machine). Thus, a principle can be formulated according to which people should retain the right to decide what decisions to make: exercise their freedom of choice where necessary and cede it in cases where overriding reasons such as effectiveness may outweigh the loss of control in making decisions[4].

Oversight may be achieved through governance mechanisms such as a human-in-the-loop (**HITL**), human-on-the-loop (**HOTL**), or human-in-command (**HIC**) approach. HITL refers to the capability

[3] According to AI HLEG: ***Accuracy*** *pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models. A* ***reliable AI system*** *is one that works properly with a range of inputs and in a range of situations.* ***Reproducibility*** *describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions*. See: AI HLEG, p. 17.

[4] L. Floridi, J. Cowls, *A Unified Framework,* p. 7.

for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system[5].

3.5 Fairness and justice

The development of AI should promote fairness and seek to eliminate all kinds of discrimination. It should also contribute to equal access to the benefits of AI technology. A serious threat to these goals is the risk of bias in the datasets used to train AI systems.

Justice also relates to using AI to remedy past mistakes, such as eliminating unfair discrimination, promoting diversity and preventing new threats to justice.

Fair and just AI provides **diversity throughout the entire life cycle of an AI system**. It requires the involvement of all stakeholders throughout the process and ensures equal access through inclusive design processes and equal treatment.

Where possible, identifiable and **discriminatory prejudices should be removed at the data collection stage**. Also at the design stage of AI systems, unfair bias must be counteracted. The implementation of data supervision and verification processes as well as decisions based on them is necessary at the testing stage, but also during the system operation. It is also important to hire people from different backgrounds, cultures and disciplines, which can provide a diversity of opinion.

AI systems should be **user-oriented**. The accessibility of this technology to people with disabilities, who are present in all social groups, is of particular importance.

3.6 Transparency, Explainability, Accountability

All documents concerning ethical AI refer to the need to understand and hold to account the decision-making processes of AI. Different terms express this principle: "transparency", "explainability", "explicability", "accountability" and "intelligibility". Each of these principles captures something seemingly novel about AI: that its workings are often invisible or unintelligible to the most expert observers. The addition of the principle of "explainability" or "explicability" incorporating both the epistemological sense of 'intelligibility' (as an answer to the

[5] AI HLEG, p. 16.

question '**how does it work**?') and in the ethical sense of 'accountability' (as an answer to the question '**who is responsible for the way it works?**'), is the crucial piece of the AI ethics[6].

AI HLEG identified three elements of transparency[7]:

- **Traceability:** The datasets and processes that lead to AI system decisions, including data collection and tagging, as well as the algorithms used, should be documented to the best possible standard to enable traceability and increase transparency. This also applies to decisions made by the AI system. This makes it possible to identify the reasons why the AI decision was wrong, which in turn can help prevent future mistakes. Traceability facilitates control of AI system.

- **Explainability:** it relates to the ability to explain both the technical processes of an AI system and related human decisions (e.g. areas of application of the system). The technical explainability requires that decisions made by the AI system can be understood and tracked by humans. Moreover, trade-offs may be necessary between improving the explainability of the system (which may reduce its accuracy) and increasing its accuracy (at the expense of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand an adequate explanation of the decision-making process of the AI system. Such explanation should be timely and adapted to the expertise of the interested party (e.g. layperson, regulatory authority or researcher).

- **Communication:** people have a right to be told that they are interacting with the AI system. This means that AI systems must be identifiable as such. Moreover, if necessary, it should be possible to decide against this interaction in favour of interpersonal interaction, to ensure the respect of fundamental rights. In addition, AI practitioners or end-users should be informed about the capabilities and limitations of the AI system in a manner appropriate to the use case.

The accountability requirement complements the above and is closely related to the principle of fairness and justice. This requires mechanisms to ensure accountability of AI systems and their outcomes, both before and after development, implementation and use.

[6] L. Floridi, J. Cowls, *A Unified Framework*, p. 8.
[7] AI HLEG, p. 18.

4 Conclusion and recommendations

1) Pilots under MAS4AI using AI should ensure that the design, implementation and use of AI are carried out in accordance with the requirements of ethical Artificial Intelligence. They also must respect all applicable laws and regulations and be robustness - both from a technical perspective while taking into account its social environment[8].

2) AI systems implemented as part of pilots must meet the requirements of the European approach to AI in terms of:

- **Trustworthy AI:** AI that is reliable;
- **Legal AI:** AI that conforms to legal requirement;
- **Ethical AI:** AI that is ethical.

3) EU human-centric approach to AI systems assumes also that these systems must **support human autonomy** and decision making, enabling users to make informed autonomous decision.

4) Each pilot should use an ***"Ethics by Design"* approach** (which can be incorporated into any design methodology). This approach assumes each time when designing AI solutions, taking into account the following "steps":

- specification of objectives (what the system is for and what it will do);
- specification of requirements (what do we need to build it - tools, processes, organisation, etc);
- data collection & preparation (data must be collected, verified, cleaned, integrated);
- detailed design and development (coding);
- testing.

5) AI HLEG translated the Trustworthy AI requirements described by it into a detailed **list of assessments**, taking into account the feedback from the six-month pilot process in the European AI community, and developed a prototype web tool that practically guides developers and AI implementers through accessible and dynamic checklist (**ALTAI**[9]). **It is recommended that pilot**

[8] High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines For Trustworthy AI*, 8 April 2019, p. 5, https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.

[9] AI HLEG, Assessment List for Trustworthy Artificial Intelligence, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

implementations use this tool for self-assessment of their systems using AI. It is available at: https://altai.insight-centre.org/

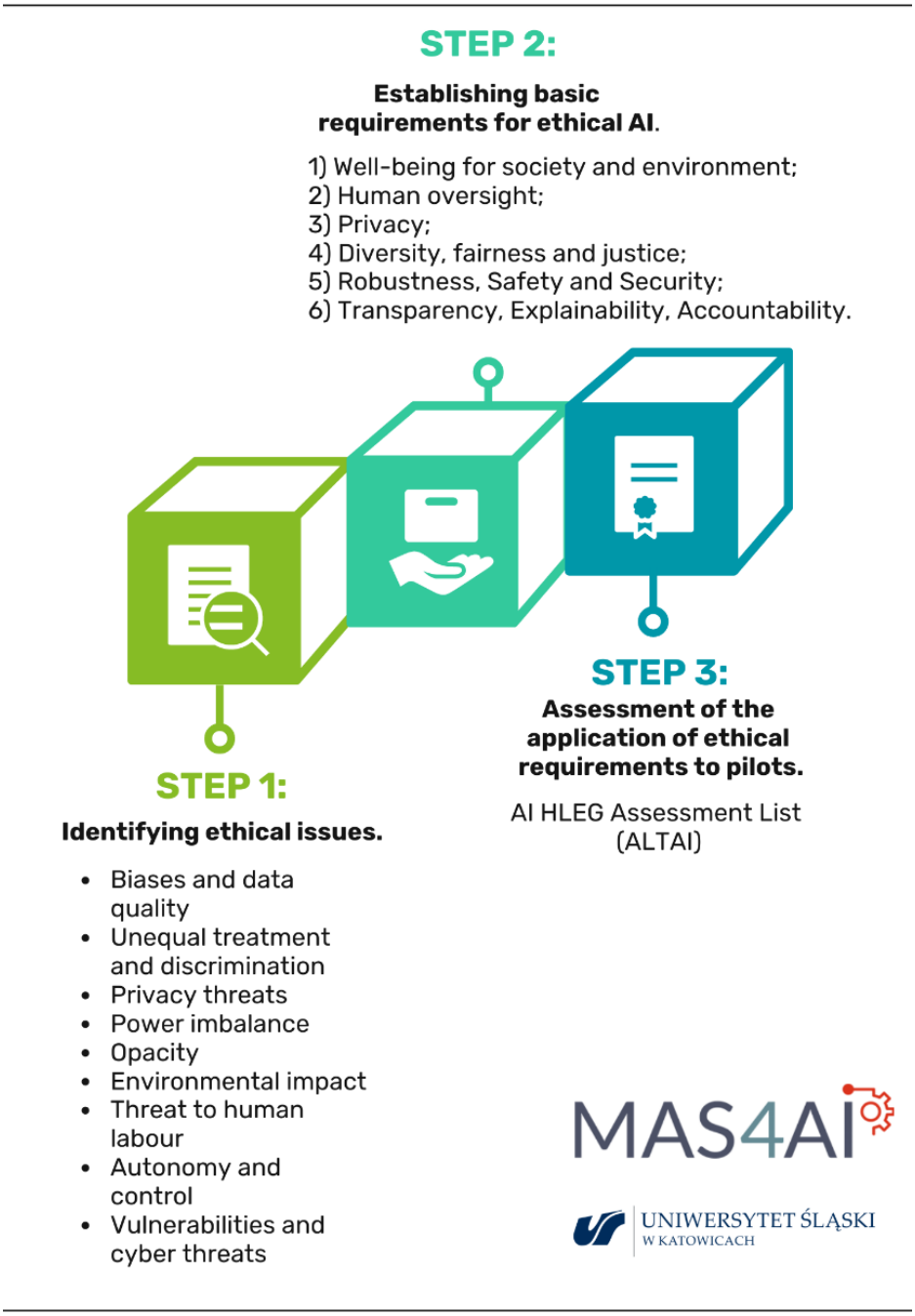

Figure 1: Infographic showing list of assessments

6) The involvement of an **ethics advisor/mentor** with appropriate expertise in ethics of new and emerging technologies is highly recommended for pilots which may raise significant ethics risks.

7) The other recommendation for team members in MAS4AI pilots is to **follow EU initiative on ethical & legal AI**:

- **European AI Alliance –** a forum engaged in a broad and open discussion of all aspects of Artificial Intelligence development and its impact, available on: https://digital-strategy.ec.europa.eu/en/policies/european-ai-alliance; https://futurium.ec.europa.eu/en/european-ai-alliance

- **AI4EU Platform –** first European Artificial Intelligence On-Demand Platform and Ecosystem with the support of the European Commission under the H2020 programme. It is available at: https://www.ai4eu.eu, https://www.ai4europe.eu/. It is worth also to take part in the consultation which will be available soon on the website: http://consultationai4eu.eu/. The survey aims to gather the views of European experts on Trustworthy AI, its implementation and governance.

- **InTouchAI.eu –** a new initiative to promote the European Commission vision on sustainable and trustworthy AI at global level. InTouchAI.eu has been launched on the initiative of the European Commission's Service for Foreign Policy Instruments (FPI), the Directorate General for Communications Networks, Content and Technology (DG CONNECT), in collaboration with the European External Action Services (EEAS). The specific objectives of the project are to support the EC to:
    - develop responsible leadership in global discussions around AI;
    - create the conditions for the uptake of **policies and good practices and standards** that ensure an appropriate ethical and legal framework on AI;
    - improve public awareness of the challenges and opportunities associated with AI.

More at: https://www.linkedin.com/company/intouchai-eu-international-outreach-for-a-human-centric-approach-to-artificial-intelligence/

- **Globalpolicy.AI –** an online platform developed through ongoing co-operation between intergovernmental organisations with complementary mandates on AI. It aims to help policy makers and the public navigate the international AI governance landscape and access the necessary knowledge, tools, data, and best practices to inform AI policy development. This is

achieved through co-operation between intergovernmental organisations that are working to promote the responsible development and use of trustworthy AI in accordance with human rights and democratic values. Available at: https://globalpolicy.ai/en/

5 References

1) Arrieta A., Diaz-Rodrigues N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garciag S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, https://arxiv.org/abs/1910.10045;

2) Council of Europe study DGI(2019)05*: A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework;*

3) Cowls J., Floridi L., and Taddeo M., *The challenges and opportunities of ethical AI,* https://digitransglasgow.github.io/ArtificiallyIntelligent/contributions/04_Alan_Turing_Institute.html

4) Cowls J., King T. C., Taddeo M. and Floridi L., *Designing AI for Social Good: Seven Essential Factors*, http://ssrn.com/abstract=3388669

5) Delcker J., *Europe's silver bullet in global AI battle: Ethics*, Politico (March 2018), https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/

6) European Group on Ethics in Science and New Technologies, *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems* (March 2018), https://ec.europa.eu/info/news/ethics-artificial-intelligence-statement-ege-released-2018-apr-24_en

7) Floridi L., *What the Near Future of Artificial Intelligence Could Be*, Philosophy & Technology 32(1): 1-15, https://doi.org/10.1007/s13347-019-00345-y

8) Floridi L., *Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical*, Philosophy & Technology 32(2): 185-193. https://doi.org/10.1007/s13347-019-00354-x

9) HLEGAI - High Level Expert Group on Artificial Intelligence, European Commission, *Ethics Guidelines for Trustworthy AI* (April 2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

10) House of Lords Artificial Intelligence Committee, *AI in the UK: ready, willing and able* (April 2018), https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

11) Khaleghi B., *The How of Explainable AI: Post modellingExplainability,* https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f

12) The IEEE Initiative on Ethics of Autonomous and Intelligent Systems (2017), *Ethically Aligned Design, v2*, https://ethicsinaction.ieee.org

13) Jezard A., China is now home to the world's most valuable AI start-up, World Economic Forum (April 2018), https://www.weforum.org/agenda/2018/04/chart-of-the-day-china-now-has-the-worlds-most-valuable-ai-startup/

14) Lee K., and Triolo, P., *China's Artificial Intelligence Revolution: Understanding Beijing's Structural Advantages*, Eurasian Group, (December 2017), https://www.eurasiagroup.net/live-post/ai-in-china-cutting-through-the-hype

15) McCarthy J., Minsky M. L., Rochester N. and Shannon C. E., *A proposal for the Dartmouth summer research project on artificial intelligence*, August 31, 1955. AI Magazine 27 (4):12. https://doi.org/10.1609/aimag.v27i4.1904

16) *Montreal Declaration for a Responsible Development of Artificial Intelligence*, announced at the conclusion of the Forum on the Socially Responsible Development of AI (November 2017), https://www.montrealdeclaration-responsibleai.com/the-declaration

17) Morley J., Floridi L., Kinsey L., & Elhalal A., *From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices,* http://arxiv.org/abs/1905.06876

18) OECD, *Recommendation of the Council on Artificial Intelligence* (2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

19) Partnership on AI (2018). Tenets. https://www.partnershiponai.org/tenets/

20) Samuel A. L. (1960). *Some Moral and Technical Consequences of Automation—A Refutation*, Science 132 (3429), 741-742. https://doi.org/10.1126/science.132.3429.741

21) Sileno G., Boer A., Van Engers T., *The Role of Normware in Trusthworthy and Explainable AI*, https://arxiv.org/abs/1812.02471

22) UNESCO, *Recommendation On The Ethics Of Artificial Intelligence*, May 2020.