Self-assessment for Human-Centred Artificial Intelligence (AI) Master's

Version 1.0 HCAIM, BME

1. Background

This document concerns MSc theses dealing with the design, development, deployment and/or use of AI-based systems or techniques.[1]

> ***Examples.*** *The uses of AI in military environments could include: decision and planning support, collaborative combat, cybersecurity and digital influence, logistics and operational, robotics and autonomy, support services and target identification and engaging.*

The manner in which an AI solution is deployed or used may change the ethical characteristics of the system. It is therefore important to ensure ethics compliance even in cases where your thesis does not develop itself an AI based system/ technique.

A proposal for a regulation laying down harmonised rules on AI (Artificial Intelligence Act) is currently pending adoption by the EU legislator. This regulation, when it enters into force, will have effect on all AI-related activities. Before its adoption and entry into force, we strongly encourage the use of the Assessment List for Trustworthy Artificial Intelligence (ALTAI) to develop procedures to detect, assess the level and address potential risks.

2. How to address the issues

Your activities must comply with ethics, notably:

- highest ethical standards;
- applicable international, EU and national law (in particular, the principles and values enshrined in the EU Charter of Fundamental rights and the EU Treaties).

This requires specific ethically-focused approach during the design, development, deployment, and/or use of AI-based solutions.

Any use of AI systems or techniques should be clearly described in the thesis and you must demonstrate their technical robustness and safety (they must be dependable and resilient to changes).

The approach must be built upon the following key prerequisites for ethically sound AI systems[2]:

Human agency and oversight — AI systems must support human autonomy and decision-making, enabling users to make informed autonomous decisions regarding the AI systems. This is particularly relevant for AI systems that can affect human behaviour by guiding, influencing or supporting humans in decision-making processes *(e.g. recommendation systems, predictive algorithms, disease diagnosing tools)*. The right to human agency should be safeguarded by setting up appropriate oversight mechanisms to prevent possible adverse effects and uphold human autonomy.

- AI systems must not subordinate, coerce, deceive or manipulate people, and should not create attachment or stimulate addiction.
- The development of lethal autonomous weapons without the possibility of meaningful human control over selection and engagement decisions when carrying out strikes against humans are prohibited.[3]

Privacy and data governance — AI systems must guarantee privacy and data protection throughout the system's lifecycle. The principles of privacy by design and by default must be taken into account in the process of designing, developing, selecting and using AI. The quality, integrity and security of data should be rigorously checked and adequately managed. Data minimisation and data protection should never be leveraged to hide or obscure bias, and these should be addressed without harming privacy rights.

Transparency — All data sets and processes associated with AI decisions must be well communicated and appropriately documented. AI systems must be explainable and open in the communication about their limitations. The principle of transparency is closely linked to the principles of tractability and explicability and facilitates the implementation of human agency, data governance and human oversight. It includes all elements relevant to an AI system *(e.g. the data, the system and the processes by which it is designed, deployed and operated)*.

Fairness, diversity and non-discrimination — Best possible efforts should be made to avoid unfair bias *(e.g. stemming from the used data sets or the ways the AI is developed)*. AI systems should be user-centric and whenever relevant, designed to be usable by different types of end-users with different abilities. AI systems should avoid functional bias by offering the same level of functionality and benefits to end-users with different abilities, beliefs, preferences and in-

[1] This assessment is based on EU Grants – How to complete your ethics self-assessment: V2.0 – 13.07.2021

terests, to the extent possible. Inclusion and diversity must be enabled during the entire life cycle of the AI system. Ensure objectivity and inclusiveness of the developed systems/approaches.

Societal and environmental well-being — The impact of the developed and/or used AI system/technique on the individual, society and environment must be carefully evaluated and any possible risk of harm must be avoided. Increased vigilance is needed for solutions that may potentially have significant negative social or environmental impact. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, for instance the Sustainable Development Goals. Overall, AI should be used to bring positive transformative changes to the society, environment or the economy. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals; they must not undermine democratic processes, human deliberation or democratic voting systems or pose a systemic threat to society at large.

> ***Examples of social impact:*** *negative impact on human rights, democratic processes, functioning of media and mass communication, labour and labour market; educational choices; consumer interests and consumer protection, social cohesion and social exclusion, cultural diversity and cultural heritage, international co-operation, mass surveillance.*

Accountability — Requires that the actors involved in the development or operation take responsibility for the way that these applications function and for the resulting consequences. Accountability requires and presupposes certain levels of transparency as well as oversight. To be held to account, developers or operators of AI systems must be able to explain how and why a system exhibits particular characteristics or results in certain outcomes.

This implies that, amongst others, the developed/used AI solutions must:

- ensure that people are aware they are interacting with an AI system and are informed (in a language and terms understandable by all) about its abilities, limitations, risks and benefits – the manner in which this is done must be described in the thesis;
    - The manner in which information is provided should not depend on particular educational backgrounds, technical knowledge, or other skills which cannot be assumed of all people.
- prevent possible limitations on human rights and freedoms *(e.g. freedom of expression, access to information, freedom of movement etc.);*
- not be designed in a way that may lead to objectification, dehumanization, subordination, discrimination, stereotyping, coercion, manipulation of people or creation of attachment or addiction;
- be able to demonstrate compliance with the principles of data minimisation and privacy by design and by default when processing personal data – the principles of lawfulness, transparency and fairness of the data processing must be respected at all times; for more information, please consult the Guidance on ethics and data protection in research projects;
- must be designed in a way to avoid bias in both input data and algorithm design – the system should be able to prevent potential discrimination, stigmatisation or any other adverse effects on the individual related to the use of the developed/deployed AI system/technique – the manner in which this is done must be described in your thesis;
- must address the potential impact on the individual, society or the environment. An evaluation of the potential negative individual, societal and/or environmental impacts must be carried out and be included in the thesis along with the measures to be set in place to mitigate any potential adverse effect;
    - The ethics risk assessment and risk mitigation measures must cover the design, development, deployment and post-deployment phases.
- must not reduce the safety and wellbeing of the individuals. Whenever relevant, the safety of the developed/used systems must be demonstrated in the thesis;
- should be developed in a way that enables human oversight (human-in-theloop, human-on-the-loop, human-in-command), traceability and auditability – whenever possible, explanation on how decisions are taken by the developed/used AI along with the logic behind it should be provided.

For further detailed requirements, please consult the Assessment List for Trustworthy Artificial Intelligence (ALTAI).

[2] As identified by the Independent High Level Expert Group on AI set up by the European Commission in the Ethics guidelines for trustworthy AI. See also: Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

[3] Article 10(6) of EDF Regulation 2021/697.

At the design and development stage, the implementation of the key requirements for ethically sound AI systems can be ensured by adopting the 'ethics by design' approach. The latter is aimed at preventing ethics issues from occurring by integrating ethical values based requirements into the design of the developed/used AI solution. The ethics by design approach will greatly facilitate your ethics compliance. For more information, please consult Guidelines on ethics by design for AI.

Some types of objectives, methodologies, system architecture or design may be inherently problematic (due to serious ethical non-compliance). This is the case for instance for AI systems that risk to:

- limit human rights, subordinate, deceive or manipulate people, violate bodily or mental integrity, create attachment or addiction, or hide the fact people are interacting with an AI system;
- cause people to be disadvantaged socially or politically, reduce the power that they have over their lives, or result in discrimination, either by the system, or by the way it will be used;
- cause people to suffer physical, psychological or financial harm, cause environmental damage, or significantly damage social processes and institutions *(for example, by contributing to misinformation of the public)*.

3. Ethics issues checklist

Activity	Y/N	Information to be provided in the thesis	Documents to be provided as appendix
Does this activity involve the development, deployment and/or use of Artificial Intelligence-based systems?		1) Explanation as to how the participants and/or end-users will be informed about: • their interaction with an AI system/ technology (if relevant); • the abilities, limitations, risks and benefits of the proposed AI system/technique; • the manner in which decisions are taken and the logic behind them (if relevant). 2) Details on the measures taken to avoid bias in input data and algorithm design. 3) Explanation as to how the respect to fundamental human rights and freedoms (e.g. human autonomy, privacy and data protection) will be ensured. 4) Detailed explanation on the potential ethics risks and the risk mitigation measures.	1) Detailed risk assessment accompanied by a risk mitigation plan (if relevant). These must cover the development, deployment and postdeployment phases. 2) Copies of ethics approvals (if relevant).
Could the AI based system/technique potentially stigmatise or discriminate against people *(e.g. based on sex, race, ethnic or social origin, age, genetic features, disability, sexual orientation, language, religion or belief, membership to a political group, or membership to a national minority)?*		1) Detailed explanation of the measures set in place to avoid potential bias, discrimination and stigmatisation.	
Does the AI system/technique interact, replace or influence human decision-making processes *(e.g. issues affecting human life, health, well-being or human rights, or economic, social or political decisions)*?		1) Detailed explanation on how humans will maintain meaningful control over the most important aspects of the decision-making process. 2) Explanation on how the presence/role of the AI will be made clear and explicit to the affected individuals.	1) Information sheets / Template Informed consent forms (if relevant).

Activity		Y/N	Information to be provided in the thesis	Documents to be provided as appendix
Does the AI system/technique have the potential to lead to negative social *(e.g. on democracy, media, labour market, freedoms, educational choices, mass surveillance)* ***and/or environmental impacts either through intended applications or plausible alternative uses?***			1) Justification of the need for developing/using this particular technology. 2) Assessment of the ethics risks and detailed description of the measures set in place to mitigate the potential negative impacts during the research, development, deployment and postdeployment phase.	1) For serious and/or complex cases: Algorithmic impact assessment/human right assessment. These must cover the development, deployment and postdeployment phases.
Does this activity involve the use of AI in a weapon system?				
If YES	I***s it possible to establish which specific function/functions are automated/autonomous in the weapon system?***		1) Justification for the need. 2) Detailed explanation on how humans will maintain meaningful control.	1) Detailed overview of the automated functions.
	If the weapon system has AI-enabled functions, could these functions render the weapon system indiscriminate?		1) Justification for the need. 2) Detailed explanation on how humans will maintain meaningful control.	1) Description of the automated navigation and its ability to discriminate targets.
	Does the design include the possibility of an autonomous mode for selfprotection? If yes, can the system reliably distinguish between targets (threats) and non-targets?		1) Justification for the need. 2) Detailed explanation on how humans will maintain meaningful control.	1) Detailed explanation on how the potential ethics algorithmic assessment will work.
Does the AI to be developed/used in the project raise any other ethical issues not covered by the questions above *(e.g., subliminal, covert or deceptive AI, AI that is used to stimulate addictive behaviours, lifelike humanoid robots, etc.)?*			1) Detailed explanation on how the potential ethics issues will be addressed and the measures set in place to mitigate ethics risks.	1) Detailed risk assessment accompanied by a risk mitigation plan. These must cover the development, deployment and postdeployment phases.

In case it is not possible to identify the potential risks related to the AI system/techniques at this stage, describe the procedure you intend to use to detect, assess and address potential ethics issues (or explain why such a procedure is not needed).

4. Background documents & further reading

Artificial intelligence

1. Proposal for an EU Regulation on a European approach for Artificial Intelligence
2. Ethics guidelines for trustworthy AI, Independent High Level Expert Group on AI
3. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment
4. Guidelines on ethics by design/operational use for Artificial Intelligence
5. EU White Paper on Artificial intelligence

Defence

1. United Nations, Convention on Certain Conventional Weapons, Group of Governmental Experts, Lethal Autonomous Weapon Systems, CCW GGE LAWS 11 guiding principles

Ethical use of generative AI in academic writing

1. The ethics of using AI in research and scientific writing, Paperpal. November 16, 2022
2. Guidance on the Appropriate Use of Generative Artificial Intelligence in Graduate Theses, University of Toronto, October 12, 2023