AI and Ethics issues in DOME

AI and risk assessment – Legal approach to AI – related risks

The analysis, assessment, and management of AI in DOME (including NLP, machine learning, and Ethics-by-design in AI) are based on the guidance provided by the Ethics Guidelines for Trustworthy AI and the Ethics By Design and Ethics of Use Approaches for Artificial Intelligence document of the EU Commission.
As clearly stated in its introduction, the Ethics By Design and Ethics of Use Approaches for Artificial Intelligence note concerns all research activities involving the development or/and use of artificial intelligence (AI)-based systems or techniques, including robotics. It builds on the work of the Independent High-Level Expert Group on AI and their mentioned Ethics Guidelines for Trustworthy AI as well as on the results of the EU-funded SHERPA and SIENNA projects.
The central approach of this note is based on Ethics by Design (see Subsection 1.3). The aim of Ethics by Design is to incorporate ethical principles into the development process allowing that ethical issues are addressed as early as possible and followed up closely during project activities. It explicitly identifies concrete tasks which can be taken and can be applied to any development methodology. However, the advised approach should be tailored to the type of activity being proposed keeping also in mind that ethics risks can be different during the research/design phase and the deployment or implementation phase.
Furthermore, from a practical viewpoint, the principles and methodology of the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment and of the HRESIA (Human Rights, Ethical, and Social Impact Assessment) model and approach will be followed.
The European AI strategy places trust as a prerequisite to ensure a human-centric approach to AI, considering AI not an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being.
In 2019, the European Commission created a High-Level Expert Group on Artificial Intelligence (AI HLEG), comprising representatives from academia, civil society, and industry to provide recommendations on future related policy development and on ethical, legal and societal issues related to AI, including socio-economic challenges. In April 2019, the AI HLEG presented the Ethics Guidelines for Trustworthy AI. According to that document, Trustworthy AI has three components, which should be met throughout the system's entire life cycle:
(1) It should be lawful, complying with all applicable laws and regulations,
(2) It should be ethical, ensuring adherence to ethical principles and values, and
(3) It should be robust; both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.
Regarding the respect of indivisible rights set out in international human rights law, the EU Treaties and the EU Charter, the following families of fundamental rights are particularly apt to cover AI systems :
1. Respect for human dignity. Human dignity encompasses the idea that every human being possesses an “intrinsic worth”, which should never be diminished, compromised or repressed by others – nor by new technologies like AI systems. In this context, respect for human dignity entails that all people are treated with respect due to them as moral subjects, rather than merely as objects to be sifted, sorted, scored, herded, conditioned or manipulated. AI systems should hence be developed in a manner that respects, serves and protects humans’ physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs.
2. Freedom of the individual. Human beings should remain free to make life decisions for themselves. This entails freedom from sovereign intrusion, but also requires intervention from government and non-governmental organisations to ensure that individuals or people at risk of exclusion have equal access to AI’s benefits and opportunities. In an AI context, freedom of the individual for instance requires mitigation of (in)direct illegitimate coercion, threats to mental autonomy and mental health, unjustified surveillance, deception and unfair manipulation. In fact, freedom of the individual means a commitment to enabling individuals to wield even higher control over their lives, including (among other rights) protection of the freedom to conduct a business, the freedom of the arts and science, freedom of expression, the right to private life and privacy, and freedom of assembly and association.
3. Respect for democracy, justice and the rule of law. All governmental power in constitutional democracies must

be legally authorised and limited by law. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation or democratic voting systems. AI systems must also embed a commitment to ensure that they do not operate in ways that undermine the foundational commitments upon which the rule of law is founded, mandatory laws and regulation, and to ensure due process and equality before the law.
4. Equality, non-discrimination and solidarity - including the rights of persons at risk of exclusion. Equal respect for the moral worth and dignity of all human beings must be ensured. This goes beyond non-discrimination, which tolerates the drawing of distinctions between dissimilar situations based on objective justifications. In an AI context, equality entails that the system's operations cannot generate unfairly biassed outputs (e.g. the data used to train AI systems should be as inclusive as possible, representing different population groups). This also requires adequate respect for potentially vulnerable persons and groups, such as workers, women, persons with disabilities, ethnic minorities, children, consumers or others at risk of exclusion.
5. Citizens' rights. Citizens benefit from a wide array of rights, including the right to vote, the right to good administration or access to public documents, and the right to petition the administration. AI systems offer substantial potential to improve the scale and efficiency of government in the provision of public goods and services to society. At the same time, citizens' rights could also be negatively impacted by AI systems and should be safeguarded. When the term "citizens' rights" is used here, this is not to deny or neglect the rights of third-country nationals and irregular (or illegal) persons in the EU who also have rights under international law, and – therefore – in the area of AI systems.

On the above mentioned groups of fundamental rights mentioned in the document, are rooted the main Ethical principles relevant in the Context of AI systems. Many of these are to a large extent already reflected in existing legal requirements for which mandatory compliance is required and hence also fall within the scope of lawful AI, which is Trustworthy AI's first component.

1. Respect for human autonomy

The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work.

2. Prevention of harm

AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. They must be technically robust and it should be ensured that they are not open to malicious use. Vulnerable persons should receive greater attention and be included in the development, deployment and use of AI systems. Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. Preventing harm also entails consideration of the natural environment and all living beings.

3. Fairness

The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives.31 The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.32 In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

4. Explicability

Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent

possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.
Based on the abovementioned fundamental rights and ethical principles, the Ethics Guidelines for Trustworthy AI set out seven key requirements that AI systems should meet in order to be trustworthy. The list of requirements is non-exhaustive, and it includes systemic, individual and societal aspects :
1. Human agency and oversight (including fundamental rights, human agency and human oversight): AI systems should support human autonomy and decision making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights and allow for human oversight. In situations where AI may negatively affect fundamental rights, a fundamental rights impact assessment should be undertaken. Users should be able to make informed autonomous decisions regarding AI systems. Human 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 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.'
2. Technical robustness and safety (including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility): A crucial component of achieving trustworthy AI is technical robustness, which is closely linked to the principle of prevention of harm. AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, e.g., by hacking. AI systems should have safeguards that enable a fallback plan in case of problems. Moreover, AI requires a high level of accuracy, which 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. It is also critical that the results of AI systems are reproducible, as well as reliable.
3. Privacy and data governance (including respect for privacy, quality and integrity of data, and access to data): AI systems must guarantee privacy and data protection throughout a system's entire lifecycle. The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. Such issues need to be addressed prior to training any given data set. In addition, the integrity of the data must be ensured. In any organisation that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place.
4. Transparency (including traceability, explainability and communication): The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability. Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g., application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. The AI system's capabilities and limitations should be communicated to AI practitioners or end-users.
5. Diversity, non-discrimination and fairness (including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation): Trustworthy AI must enable inclusion and diversity throughout the entire AI system's life cycle, ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with the principle of fairness. Trustworthy AI requires avoidance of unfair bias, accessibility and universal design and stakeholder participation. Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. This could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. The way in which AI systems are developed (e.g., algorithms' programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner. Moreover, hiring from diverse backgrounds, cultures and disciplines can ensure

diversity of opinions and should be encouraged. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards. It is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle.
6. Societal and environmental well-being (including sustainability and environmental friendliness, social impact, society and democracy): Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as the UN's Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations. The effects of AI on people's physical and mental wellbeing, society and democracy must be carefully monitored and considered.
7. Accountability (including auditability, minimisation and reporting of negative impact, trade-offs and redress): Trustworthy AI necessitates mechanisms to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use. It requires (a) auditability, which entails the enablement of the assessment of algorithms, data and design processes, (b) minimisation and reporting of negative impacts through impact assessments used (e.g., red teaming or forms of algorithmic impact assessment) both prior to and during the development, deployment and use of AI systems, (c) trade-offs should be addressed in a rational and methodological manner within the state of the art, (d) when unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.
In July 2020, the AI HLEG released a self-assessment list to help AI developers assess their tools against the Ethics Guidelines for Trustworthy AI . This list will guide all AI developers through DOME.
Finally, the Institute of Electrical and Electronic Engineers (IEEE), a global professional organisation working towards technology standards for human benefit, has a global Initiative on the Ethics of Autonomous and Intelligent Systems. In 2019, the IEEE released Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems outlining a number of general principles to guide technology developers in the design and implementation of Autonomous and Intelligent Systems (A/IS) . These principles, that will be taken into consideration throughout DOME, are the following:
1. Human Rights. A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. Well-being. A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. Data Agency. A/IS 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.
4. Effectiveness. A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. Transparency. The basis of a particular A/IS decision should always be discoverable.
6. Accountability. A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. Awareness of Misuse. A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. Competence. A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

Revision #1
Created 29 May 2024 13:10:31
Updated 29 May 2024 13:11:25