Responsible Research Brief

HIGHLIGHTS

DARLENE is an EU-funded Horizon 2020 project investigating how AI-based Augmented Reality tools can be used by law enforcement agents to support the visual understanding of complex real-world scenes and improve their decision-making processes in high-pressure scenarios.

This brief provides guidance to researchers involved in innovation projects. It shares some practical recommendations on how to:

1. promote responsible and ethical development when designing new technologies
2. Transform from theory to practice principles of responsible development.

The risk levels are: unacceptable, high, low or minimal. AI practices that contravene EU values or infringe upon fundamental rights, by, for example, manipulating people or exploiting their vulnerabilities, **are prohibited** due to the unacceptable level of risk they pose. AI practices that create a **considerable risk** to health and safety, environment, fundamental rights and rule of law are categorised as high risk and are only allowed to be placed on the EU market when having demonstrated compliance with certain mandatory requirements for trustworthy AI through a conformity assessment. Systems that might pose a **low or limited risk** to EU values and fundamental rights by (i) interacting with humans, (ii) detecting emotions or determining association with (social) categories based on biometric data, or (iii) generating or manipulating content are required to comply with specific transparency obligations.

BACKGROUND

In April 2021, the European Commission published a proposal for the EU AI Act, for which the European Parliament and the Council reached a provisional agreement on 9 December 2023. The AI Act adopts a risk-based approach, where various types of AI systems are regulated according to the **potential risks** they pose to individuals or society as a whole.

Despite assigning different requirements to different risk levels, the AI Act, under Title IX, encourages providers of non-high-risk AI systems to voluntarily apply the requirements for high-risk AI systems and adopt relevant code of conducts.

THE CHALLENGE

Achieving a clear understanding of whether an AI-based tool like DARLENE will be classified as high risk is challenging until a relatively advanced stage of technology development. The level of risk might also vary depending on the context in which the technology is used.

We suggest to incorporate legal and ethical assessments into the technology design process early on.

This will allow you to:

- Improve human centric design;
- Better understand the risks involved;
- Early plan mitigation strategies;
- Be ready to demonstrate compliance if, in the end, the technological functionalities or the context in which the technology will be used place it in the high risk category.

Our recommendations intend to help researchers to better incorporate a socio-technical approach to technology development projects, including better understanding the AI Act requirements and how they can be integrated into the research process.

RECOMMENDATIONS

These recommendations translate the AI Act requirements into some actionable steps that researchers can take during the development process to embed legal and ethical assessments into the design of the technology.

We emphasise the importance of **co-creation**. Involving end-users, technical experts, social scientists, representatives from civil society and more in the technology development process is crucial to create responsible technologies.

1. Manage risks

The AI Act requires to establish, implement, document and maintain a risk management system. This risk management system needs to be an iterative process that runs through the entire lifecycle of the AI system.

Identify and evaluate possible risks

As the technology evolves, risks might also change.

Throughout the design process, systematically identify and assess the risks posed by the technology to various groups of people who may be directly or indirectly affected by it.
To do so, map the different groups of people that might be directly or indirectly affected by the technology and for each group, identify both benefits and risks.

This should be a recurring practice during the development process, possibly involving users, the technical team, social scientists, representatives from civil society and citizens.

Adopt risk management measures

Identify, document and adopt measures to avoid or mitigate risks. Try to involve different types of stakeholders when planning your risk management strategy: participation is key to designing effective measures!

2. Govern your data

When developing a new system, you should use appropriate data governance and management practices.

Consider:

- Design choices;
- Data collection procedures;
- Data preparation operations including data annotation, labelling, cleaning and aggregation;
- How you formulate assumptions regarding the information that data are supposed to measure and represent;
- Assessing a priori the availability, quantity and suitability of the data;
- Examining biases;
- Identifying any possible data gaps and shortcomings and how these can be addressed.

Evaluate your training, validation and testing datasets. They should be relevant, representative, free of errors and complete.

Think about the context where the AI system is going to be used. This will help you to evaluate the datasets based on the impact that the system might have.

Plan a way to monitor and detect bias when the system will be in use.

3. Put together the technical documentation

The AI Act requires to out together technical documentation. Collecting all the information at an early stage is helpful to build a common understanding of the project activities and trajectory.

Clarify the intended purpose

Having a clear understanding of the system's purpose from the start will enable you to make more informed design decisions. It's also beneficial to periodically update the system's purpose as the project progresses. This allows you to reflect on how it has evolved, which can be helpful in predicting its future direction.

Keep track of the AI decision-making

Having a good understanding of the logic of the AI system and its algorithms as well as the reason behind the deign choices and the assumptions made is useful for several reasons:

- It supports more effective communication and collaboration between team members.
- It facilitates the adaptation and customisation of different features as the project evolves.
- It helps to identify potential ethical challenges and to address them proactively.

Document your dataset

Create datasheets and maintain them with up-to-date information, including details about the training data such as their source, the methods and techniques used for cleaning and training.

4. Keep the record

The system should be designed to record its operational events. This is a way to making the system working traceably and makes its output more understandable.

Consider including in the recording:

- The period of each use of the system;
- The identification of the person involved in dealing with the results of the system;
- The database against which the input data has been checked by the system (where applicable);

  The input data for which the search has led to a match.

5. Ensure transparency to users

AI systems should be designed to be transparent to users who should be able to interpret and use the output of the system appropriately. To this end, the instructions accompanying the system should be accessible and concise.

Involve users

When drafting the instructions and training material, involve users at every stage of the process. Their direct input and insights are important to shape content that is not only accurate and clear but also user-friendly and effective.

Involving users will make it easier to bring the training and instruction material to their level. It will improve the clarity and transparency of the content, leading to a better understanding of the technology and a more informed and responsible use of it.

6. Harness human oversight

AI systems should be designed in a way that includes mechanisms to allow human oversight when the system is in use. Humans using the technology should be able to understand the capacities and limitations of the system, monitor its working and be able to detect dysfunctions or unexpected results.

Co-create the technology

Plan participatory co-creation activities from the early development stage. This will allow for the development of more human-centred technologies that meet the needs and expectations of different user groups. This can result in a more intuitive interface and a better overall user experience.

Diversify the people involved in the development process to incorporate more views and perspectives.

Collect user requirements

Collect and filter user requirements based on their technical, legal, and ethical feasibility, and on their alignment with the project objectives. Not all ideas or suggestions may be suitable for implementation within the project scope, timeframe or resources. Rejecting ideas that cannot be implemented should be done in a thoughtful and considerate manner, to avoid that the end users involved feel rejected and disengaged. Being clear and transparent about the decision-making process from the very beginning is key to avoid their frustration and for good expectation management. The procedure should include providing constructive feedback, and technical as well as legal and ethical explanations in an accessible way to the parties involved.

7. Embed accuracy, robustness and cybersecurity

The AI system should reach a good level of accuracy, robustness and cybersecurity and perform consistently throughout its lifecycle. Be transparent about the levels of accuracy: they should be part of the training material and instructions for users.

Build systems that are resilient regarding errors and inconsistencies that might happen when the system is in use.

Embed in the system ways to avoid unauthorised third parties to alter their use and performance.

Integrate cybersecurity options that are appropriate to the levels of risk.

FINAL TAKE-AWAY

When working on innovation projects involving the development of new AI-based technologies, it is important to integrate legal and ethical considerations at a very early stage of development.

Using the AI Act requirements for high-risk systems as a base ground for legal and ethical thinking during the project development is a good way to:

- have more responsible technologies that align better with regulations and societal values;
- proactively address potential risks and challenges.

FURTHER READING

European Commission, 2021. Proposal for a Regulation of the European Parliament and of the Council Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts.

Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts

European Parliament press release, 2023. Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI

High Level Expert Group on AI. Ethics Guidelines for Trustworthy AI

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 883297.

By
Trilateral Research