SmartEdge

Semantic Low-code Programming Tools for Edge Intelligence

This project is supported by the European Union's Horizon RIA research and innovation programme under grant agreement No. 101092908

Deliverable D1.5

Ethical Guidelines

Editor D. Raggett, ERCIM.

Contributors D. Raggett, ERCIM, R. Wenning, ERCIM, F. Cugini, CNIT

Version 1

Date 15 June 2023

Distribution Public

DISCLAIMER

REVISION HISTORY

Revision	*Date*	*Responsible*	*Comment*
0.1	*10 May 2023*	*D. Raggett (ERCIM)*	*First version*
0.2	*10 June 2023*	*D. Raggett (ERCIM)*	*Updated structure*
1	*15 June 2023*	*D. Raggett (ERCIM)*	*Final Version*

LIST OF AUTHORS

Partner	*Author*
ERCIM	*D. Raggett*
ERCIM	*R. Wenning*
CNIT	*F. Cugini*

EXECUTIVE SUMMARY

This document introduces ethical guidelines for the SmartEdge project, which has received funding from the European Union's Horizon Program under Grant Agreement number 101092908.

We outline general principles applicable to the SmartEdge project use cases in regard to the administrative and technical aspects of ethical data management. We cover requirements under existing EU legislation in respect to handling personal data as well as specific considerations for applying AI at the Edge. The latter is based upon the Assessment List for Trustworthy Artificial Intelligence (ALTAI) produced by the High-Level Expert Group on AI established by the European Commission in 2018. We also discuss the ideas under discussion for the proposed EU AI Act which will classify AI systems into risk categories and mandate corresponding development and use requirements.

Note that this deliverable is designed to complement D1.1 Data Management Policies.

TABLE OF CONTENTS

1 INTRODUCTION

This document is intended as guidance on ethical considerations for the SmartEdge use cases. We start with a brief introduction to each of the use cases. Section 2 then covers ethical considerations relating to privacy and personal data. Section 3 covers ethical considerations relating to AI, safety and liability. Section 4 relates these considerations to the use cases. An updated version of this document will be provided for the second Review period.

1.1 INTRODUCTION TO THE SMARTEDGE USE CASES

SmartEdge has 5 major use cases. A common theme is the idea of swarm computing, in which a number of entities work collectively to fulfil shared goals. SmartEdge will define a common framework for entities to communicate as part of a swarm, along with the means to coordinate swarm behaviour. All use cases have security requirements in respect to physical faults and cyberattacks. Some use cases may involve commercially sensitive data, whilst others may involve handling personally identifying data. Edge processing can avoid the need to store and transmit images and video which may contain personal data, e.g., car number plates and people's faces. Instead, the camera data is dynamically transformed at the edge into structured data retaining only the information needed for the given task.

1.1.1 Use Case 1: Smart Vehicle to Vehicle

This is a virtual reality simulation of vehicles and traffic scenarios as a basis for evaluating the safeness of Advanced Driving Assistant Systems (ADAS), such as automatic emergency braking, lane departure detection, pedestrian detection, surround view, lane tracking, parking assist, driver drowsiness detection and gaze detection. This use case treats vehicles as participants in a swarm featuring vehicle to vehicle communication.

Question: will this use case exploit traffic cams as suggested by Figure 22 in D2.1? My guess is that such data will only be used indirectly to train generation of synthetic data for simulated traffic. In other words, we won't use the car registration numbers or images of people such as drivers, passengers and pedestrians.

1.1.2 Use Case 2: Smart Vehicle to Infrastructure

This focuses on Smart operation of traffic lights for controlling road junctions to optimise traffic flow. The use case features vehicles, traffic lights, and roadside sensors such as microwave radars, under-road vehicle sensors and traffic cams. Smart vehicles communicate with roadside edge-boxes to convey information on the vehicle's state, e.g., braking, signalling, current gear.

This use case appears to capture personal data and as such incur GDPR compliance. How will this use case handle personally identifying data? Does this use case provide special treatment for emergency response vehicles such as ambulances, fire trucks and police cars? Will cameras be used to detect signalling, lane changes, braking/accelerating, etc. and if so, will this involve personal data such as the vehicle number plates?

1.1.3 Use Case 3: Smart Factory: Mobile Robots

This focuses on smart factories using autonomous mobile robots (AMRs) to move workpieces from one part of a factory to another, e.g., manufacturing cells or storage areas. Rather than carrying the workpieces directly, this particular use case features mobile racks that the AMRs can tow to where the workpieces are needed. The AMRs form a swarm whose behaviour collectively optimise transport, avoiding collisions and adapting to uncertain environments, e.g., involving people moving around the factory floor, and unexpected obstacles, e.g., misplaced boxes. AMRs have multiple sensors, e.g., LiDAR, bump sensors, and cameras, along with actuators for grabbing and releasing the mobile racks. Robot arms are used to move workpieces between conveyor belts and the mobile racks.

This use case involves commercially sensitive data that must be kept confidential. No personal data is needed.

1.1.4 Use case 4: Smart Factory: Low-Code Edge Intelligence

This use case addresses applying semantic technologies to low-code approaches for dynamically adapting smart manufacturing systems to deal with new requirements and unexpected situations, e.g., clogging events. This includes support for manufacturing highly individualised products that involve variations in materials and processing steps according to the customer's order. This is enabled through the use of semantic models as part of the planning process.

This use case involves commercially sensitive data that must be kept confidential. The only personal data is that included in the customer orders.

1.1.5 Use Case 5: Smart Healthcare

This use case seeks to improve the quality of life and care for elderly residents in care homes. This is accomplished using a combination of eHealth devices, AI and personal virtual assistants, that track the health and wellbeing of each resident, and support care staff in decision making about care delivery. A distinct feature is the need for non-repudiation in respect to medical interventions, given the potential for legal challenges.

The approach involves collecting data from sensors, residents and care staff, its aggregation and analysis in near real-time, and presentation in support of decisions by care staff around medical interventions and care plans. This use case will be deployed in a controlled laboratory setting that simulates the environment of a real care home.

Using a simulated environment rather than a real one avoids the need to seek approval from ethical committees. Nevertheless, this use case will apply the data handling practises that would be needed to comply with the legal and ethical regulations that apply to real-life settings.

2 ETHICAL CONSIDERATIONS RELATING TO PERSONAL DATA

This section describes ethical and regulatory compliance considerations for handling personal data.

2.1 Sources of Ethical Values

2.1.1 Why Privacy is a Fundamental Right

For the ethics considerations, it is useful to evoke the goals against which goals and benefits will be measured against. People are often in the dark: Must this personal data be transferred, why, how, can it be under pseudonym, would a bug leak information etc. Is anonymity the ultimate goal? On the one hand, there are too many people who have *"nothing to hide"*. They give away their data and do not see the threat, which bears the question: *"Can we do without?".* On the other hand, there is a privacy administration with laws and rules. Sometimes one can barely dismiss the feeling that this administration is self-referential and thus *"Privacy"* is just the absolute object they are administrating. So why do we need privacy?

Beate Rössler from Amsterdam University opened many people's eyes with a book: "*Der Wert des Privaten (The Value of Privacy)* [1]*"* where she diligently and scientifically works out the variety of privacy concepts in philosophy, fears, claims, complaints, to the one and overarching architectural term: **Autonomy!** This is the overarching concept. It explains why we are talking about a human right.

As an anecdote, let me remind you the story behind the birth of the right to privacy, invented by Judge Louis Brandeis in 1890[2]. Brandeis had a law firm in Boston with a partner: Samuel D. Warren. At this time, the technical advances in photo cameras were such that journalists were just about to be able to carry them around. Warren was photographed walking hooked with a very nice women, who was not his wife. At those times a rather grave sinner. The photo was published in a newspaper. The consequences nearly killed the law-firm and Warren's career. Brandeis invented the right to be left alone. Today again, we have new technology, new challenges but still the same society that sanctions certain behaviours, links, contacts etc.

2.1.2 Roles and Opaqueness

Every one of us has a variety of different roles in society. There is private life, work life, social engagement, friends, sports and so on. All those roles come with different rules and expectations. A matrix perhaps that can be hardly transposed into a computer program. In all those roles, humans are rather good in displaying a certain image that may vary from role to role. If information from one role spills over to another role, the image is tainted thus resulting in a risk of decreased social success. This is the very reason we see people fighting to prevent certain information about them leaking to the public.

Spilling over can come from deliberate release of said information in one context and re-surface in another. It can come from the data trails we leave on the net, that surface in some way to people of a different role/context. The most tangible and urgent issue are people that communicate with friends on social networks and do not realise that an employer can look up such information thus contradicting the polished image one gives in the candidature for a new job. But information can not only spill over from roles. Time is also a very important axis to think along. Would you want to be confronted with the stupid things you did at age 16?

But how would we know whether somebody knows something about us that may taint our image? On the Internet and the Web, we are well aware of the effects of fear, uncertainty and doubt. In the analogue world there are two factors helping our data subject: Paper has to be handed over and the spreading is limited by the paper factor. So,

there is a wilful act in some way. The issue people fear most with the new devices and the networked society is that human senses are not able to detect what is going on. This can lead to the fear that everything done is assumed to be known. This has scientifically proven to cause disastrous effects on the self-confidence with which one dares defending minority opinions.

If opinions can't be exchanged without fear, uncertainty and doubt, the accumulated opinion will not be the same. But the unbiased opinion building is seen as a necessary pre-condition for the functioning of a democracy. And an unbiased building of opinion is only possible if people are still autonomous enough to make their own decision.

This is the reason, why dogmatically, the right to privacy is also attached to the freedom of expression in some jurisdictions. Privacy in this sense of data protection wants to preserve the freedom of expression. An individual fearing disclosure of image tainting information will not use his or her freedom of expression thus impeding the democratic process.

This means, the ultimate value on the other side is to protect the democratic process and the autonomy of decision making by individuals. Privacy and data protections are means that serve this goal and not absolute values in their own right.

2.2 Regulations and Treaties

Personal data is a regulated area in the EU. Since Kant we know that laws are a formalized morale or ethics. Laws are thus a very important and prime source for ethical values. Laws themselves have some hierarchical structure. Constitutions normally prime over simple law and Union law primes over national law, unless the rules of subsidiarity tell us otherwise. Special regulations for a specific area prime over more generic rules that may complement them. SPECIAL is a Horizon Europe project and thus the Horizon Europe Regulation is the nearest source of rules. The constitutional rules on EU level and the supranational rules give hints about the interpretation of those rules. Process wise, an assessment has to take all those rules into account if they are applicable. But even if not applicable, the rules can hint at an ethical value to follow.

2.2.1 Horizon Europe rules

Consideration number 71 of the Horizon Europe Regulation (2021/695/EU)[3] requests research and innovation activities supported by Horizon Europe to respect fundamental ethical principles. This includes the Charter of Fundamental Rights of the European Union and the conformity with legal obligations and Commission decisions to preserve and comfort the research integrity. The opinions of the European Group on Ethics in Science and New Technologies (EGE), the European Union Agency for Fundamental Rights and the European Data Protection Supervisor should be taken into account.

The relevant actual legal text is laid down in Article 19:

Article 19

1. Actions carried out under the Programme shall comply with ethical principles and relevant Union, national and international law, including the Charter and the European Convention for the Protection of Human Rights and Fundamental Freedoms and its Supplementary Protocols.

Particular attention shall be paid to the principle of proportionality, to the right to privacy, the right to the protection of personal data, the right to the physical and mental integrity of a person, the right to non-discrimination and to the need to ensure protection of the environment and high levels of human health protection.

2. Legal entities participating in an action shall provide:

(a) an ethics self-assessment identifying and detailing all the foreseeable ethics issues related to the objective, implementation and likely impact of the activities to be funded, including a confirmation of compliance with paragraph 1 and a description of how it will be ensured;

(b) a confirmation that the activities will comply with the European Code of Conduct for Research Integrity published by All European Academies and that no activities excluded from funding will be conducted;

(c) for activities carried out outside the Union, a confirmation that the same activities would have been allowed in a Member State; and

(d) for activities making use of human embryonic stem cells, as appropriate, details of licensing and control measures that shall be taken by the competent authorities of the Member States concerned as well as details of the ethics approvals that shall be obtained before the activities concerned start.

3. Proposals shall be systematically screened to identify actions which raise complex or serious ethics issues and submit them to an ethics assessment. The ethics assessment shall be carried out by the Commission unless it is delegated to the funding body. All actions involving the use of human embryonic stem cells or human embryos shall be subject to an ethics assessment. Ethics screenings and assessments shall be carried out with the support of ethics experts. The Commission and the funding bodies shall ensure the transparency of the ethics procedures without prejudice to the confidentiality of the content of those procedures.

4. Legal entities participating in an action shall obtain all approvals or other mandatory documents from the relevant national, local ethics committees or other bodies, such as data protection authorities, before the start of the relevant activities. Those documents shall be kept on file and provided to the Commission or the relevant funding body upon request.

5. If appropriate, ethics checks shall be carried out by the Commission or the relevant funding body. For serious or complex ethics issues, ethics checks shall be carried out by the Commission unless the Commission delegates this task to the funding body.

Ethics checks shall be carried out with the support of ethics experts.

6. Actions which do not fulfil the ethics requirements referred to in paragraphs 1 to 4 and are therefore not ethically acceptable, shall be rejected or terminated once the ethical unacceptability has been established.

2.2.2 Sources for Fundamental Rights

Section 1 of Article 19 points to the Charter of Fundamental Rights of the European Union. Article 7 and 8 of the Charter deal with privacy:

> Article 7 – Respect for private and family life Everyone has the right to respect for his or her private and family life, home and communications. Article 8 – Protection of personal data
>
> Everyone has the right to the protection of personal data concerning him or her.
>
> Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
>
> Compliance with these rules shall be subject to control by an independent authority.

Section 1 also mentions the European Convention on Human Rights (ECHR). The ECHR creates Privacy as a human right in Article 8:

> Article 8 – Right to respect for private and family life
>
> Everyone has the right to respect for his private and family life, his home and his correspondence.
>
> There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

Based on the general human right to privacy, the Council of Europe has also elaborated the Convention 108 for the Protection of Individuals with Regard to the Processing of Personal Data that was modernised in 2016. It is a convention creating an international law that binds the Members of the Council of Europe. It contains the principles of Privacy derived from Article 8 and applied to data protection. As the Council of Europe has reach beyond the European Union, those principles help to implement data protection in the member states of the Council of Europe. The new GDPR is certainly a full implementation of the international duties derived from Convention 108 in its newest iteration. Convention 108 is interesting because it contains a chapter on transborder data flows. With the GDPR of the EU, this loses the importance for the use cases in SPECIAL but remains an important orientqation for use cases where data flows outside the EU.

2.2.3 The General Data Protection Regulation

The GDPR is in the centre of attention of Smartedge. It determines the legal framework for the consent requirements and the transparency duties. Smartedge will cater to those requirements using agreements, technical and organisational means for compliant data processing. A detailed assessment of the use cases will follow in section 4.

2.3 The Opinions of the European Data Protection Board (EDPB)

GDPR has established the European Data Protection Board. It is the independent EU Advisory Body on Data Protection and Privacy. Its tasks are laid down in Chapter VII section 3 of the GDPR. The EDPB issues Opinions and Guidelines. Both are relevant for SmartEdge, especially the following:

1. **Guidelines 05/2020** on consent under Regulation 2016/679[5]

2. **Guidelines 01/2020** on processing personal data in the context of connected vehicles and mobility related applications [6]
3. **Guidelines 04/2019** Data Protection by Design and by Default [7]
4. **Guidelines 3/2019** on processing of personal data through video devices [8]
5. **Guidelines 2/2019** on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects [9]
6. **WP248rev.01** Guidelines on Data Protection Impact Assessment (DPIA) [10]
7. **WP 221** on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU [11]
8. **WP 216** on Anonymisation Techniques [12]
9. **WP 202** on apps on smart devices [13]
10. **WP 187** on the definition of consent is also somewhat relevant to SmartEdge. Consent is one of the possible responses to the ethical issues raised.[14]
11. **WP 185** on Geolocation services on smart mobile devices is directly relevant to two of the SmartEdge use cases [15]
12. **WP 163** on online social networking to partially reflect the feedback mechanisms and their potential sharing with other data subjects [16]
13. **WP 115** on the use of location data with a view to providing value-added services [17]

All those opinions contain important ethical decisions. They help to assess the areas where a higher risk is created and also establish very detailed requirements for the mitigation of the risk found.

2.4 The Grant Agreement

ARTICLE 14 – ETHICS AND Values (page 51)

14.1 Ethics

The action must be carried out in line with the highest ethical standards and the applicable EU, international and national law on ethical principles.

Specific ethics rules (if any) are set out in Annex 5.

14.2 Values

The beneficiaries must commit to and ensure the respect of basic EU values (such as respect for human dignity, freedom, democracy, equality, the rule of law and human rights, including the rights of minorities).

Specific rules on values (if any) are set out in Annex 5:

ANNEX 5:

ETHICS (— ARTICLE 14)

Ethics and research integrity

The beneficiaries must carry out the action in compliance with:

- ethical principles (including the highest standards of research integrity)

and

- applicable EU, international and national law, including the EU Charter of Fundamental Rights and the European Convention for the Protection of Human Rights and Fundamental Freedoms and its Supplementary Protocols.

No funding can be granted, within or outside the EU, for activities that are prohibited in all Member States. No funding can be granted in a Member State for an activity which is forbidden in that Member State.

The beneficiaries must pay particular attention to the principle of proportionality, the right to privacy, the right to the protection of personal data, the right to the physical and mental integrity of persons, the right to non-discrimination, the need to ensure protection of the environment and high levels of human health protection.

The beneficiaries must ensure that the activities under the action have an exclusive focus on civil applications.

The beneficiaries must ensure that the activities under the action do not:

- aim at human cloning for reproductive purposes
- intend to modify the genetic heritage of human beings which could make such modifications heritable (with the exception of research relating to cancer treatment of the gonads, which may be financed)
- intend to create human embryos solely for the purpose of research or for the purpose of stem cell procurement, including by means of somatic cell nuclear transfer, or
- lead to the destruction of human embryos (for example, for obtaining stem cells).

Activities involving research on human embryos or human embryonic stem cells may be carried out only if:

- they are set out in Annex 1 or
- the coordinator has obtained explicit approval (in writing) from the granting authority.

In addition, the beneficiaries must respect the fundamental principle of research integrity — as set out in the European Code of Conduct for Research Integrity.

This implies compliance with the following principles:

- reliability in ensuring the quality of research reflected in the design, the methodology, the analysis and the use of resources
- honesty in developing, undertaking, reviewing, reporting and communicating research in a transparent, fair and unbiased way
- respect for colleagues, research participants, society, ecosystems, cultural heritage and the environment
- accountability for the research from idea to publication, for its management and organisation, for training, supervision and mentoring, and for its wider impacts

and means that beneficiaries must ensure that persons carrying out research tasks follow the good research practices including ensuring, where possible, openness, reproducibility and traceability and refrain from the research integrity violations described in the Code.

Activities raising ethical issues must comply with the additional requirements formulated by the ethics panels (including after checks, reviews or audits; see Article 25).

Before starting an action task raising ethical issues, the beneficiaries must have obtained all approvals or other mandatory documents needed for implementing the task, notably from any (national or local) ethics committee or other bodies such as data protection authorities.

The documents must be kept on file and be submitted upon request by the coordinator to the granting authority. If they are not in English, they must be submitted together with an English summary, which shows that the documents cover the action tasks in question and includes the conclusions of the committee or authority concerned (if any).

VALUES (— ARTICLE 14)

Gender mainstreaming

The beneficiaries must take all measures to promote equal opportunities between men and women in the implementation of the action and, where applicable, in line with the gender equality plan. They must aim, to the extent possible, for a gender balance at all levels of personnel assigned to the action, including at supervisory and managerial level.

2.5 Mitigation Techniques

SmartEdge has a variety of mitigation techniques at their disposition to help with privacy and data protection concerns.

2.5.1 The Goals of Mitigation

In the risk assessment, the privacy assets and the dangers for data subjects have been identified. This chapter will address the possible mitigation of those risks and dangers. The mitigation focuses on the technical aspects and also includes the actions legally necessary. The analysis in some points contains suggestions that go beyond the strictly legally necessary.

The Ethics of information and communication technologies of the EGE has a rather short section on data protection issues [18]:

Individuals need sufficient control of their online data to enable them to use the Internet responsibly. Clarification concerning the conditions for the data subject's consent should therefore be provided, in order to always guarantee informed consent and ensure that the individual is fully aware that he or she is consenting to data processing and what it entails, in line with Article 8 of the EU Charter of Fundamental Rights.

The EGE hints to control by the data subject, consent, and transparency. For the European Data Protection Supervisor (EDPS), the ethical challenges centre around:

1. human dignity,
2. accountable controllers,
3. empowered users,
4. innovative privacy engineering.

Those are tangible goals for mitigation techniques, though not all of them have a direct technical implication. The overall direction is given by the first challenge: human dignity. A mitigation of a risk thus has to work to safeguard or to reinstate the human dignity of a data subject when confronted with the risks that constitute the processing of his personal

data. A clear ethical goal is thus given to the task of mitigation. The ethical challenge thus is the counterbalancing of several values against each other. And human dignity has to be respected while pursuing very important other values, like human health, the advancement of medical research, economic development, etc.

In this respect, the EDPS hints at a possibly difficult dilemma between values. In a given situation, a balance may be hard to strike. But innovative privacy engineering can alter the situation to a point where the ability to support privacy facilitates the construction of a system rather than prohibiting it. To put it with the words of the EDPS: "Technology should not dictate values and rights, but neither should their relationship be reduced to a false dichotomy".

The SmartEdge project is consistent with this view that technology should facilitate the support of data protection requirements. Ethically, SmartEdge is rather an enabler, as it brings new means to mitigate and control risks to the edge-cloud continuum. It does so by applying Linked data technology that allows to leverage insights from the SPECIAL and that otherwise would have led to the prohibition of the MOSAICrOWN project. The "Privacy by Design" approach itself does not tell people what to do in detail. But it helps in the decision making once the service design is at that level of detail.

The principle of privacy by design wants us to take into account the tensions between conflicting interests and values at a very early stage. If a new service is designed, the burdensome work of resolving those conflicts has to be done first. Is the full detailed record of the data subject's conversations really needed? Does the data collected serve the goal of the service or is it just "good to have"? The principle of data minimization is present as early as in the OECD Privacy Guidelines from 1981. Many legal texts help to resolve those conflicts.

The work in SmartEdge will support the explicit representation and consideration of data protection requirements and will facilitate a more careful analysis of the balance between utility extracted from the data and the costs associated with privacy.

2.5.2 The basic Principles

One of the oldest sources for the ethical balance of the processing of personal data are the OECD Privacy Guidelines from 1981[19]. They were revised in 2013[20]. From an ethical point of view, the basic principles are most useful for ethical considerations:

- Collection Limitation Principle,
- Data Quality Principle,
- Purpose Specification Principle,
- Use Limitation Principle,
- Security Safeguards Principle,
- Openness Principle,
- Individual Participation Principle,
- Accountability Principle.

The principles help to take into consideration the needs of data subjects. It is noteworthy that the European Union GDPR [1] has fewer principles but is known as the stricter framework. While the OECD Guidelines are a compromise of a large number of governments, including the United States, the EU GDPR tries to establish a set of strict and comprehensive rules. Comparing them goes beyond what this document could do. But it is still useful to see the principles side by side. The principles of the GDPR include:

- Lawfulness, fairness and transparency,
- Purpose limitation,
- Data minimization,
- Accuracy,
- Storage limitation,
- Integrity and confidentiality,
- Accountability.

As SmartEdge is an EU project, in case of differences, the GDPR will be the primary source for the decision over ethical conflicts.

2.5.3 Keeping Things in Context

As we have seen earlier, one of the major risks for individuals is that information given in one context reappears in another, unwanted, context. To mitigate those risks, data processing has to be made context aware. But this is a difficult task as context is a very difficult concept to implement. SmartEdge will allow to add privacy information within the edge-cloud continuum and thus take into account all environmental information. This includes protocol information, provenance information, policy information as far as they are available. Consequently, this information is present at collection time and will be remembered. Later processing now has that information to avoid contextual violations. SmartEdge is especially innovative here as it allows to filter the information already in the data stream coming from the edge.

The amount of contextual integrity also depends on the cultural context. On the Internet, the sharing of information is built into the system. The dominating US context and the fact that the US still has no comprehensive federal privacy regulation means that the free sharing of information, regardless of context, is the default. It is therefore important to take into account that a service should be very careful about what information is shared with other actors.

The vision of the digital data market is to support the realization of benefits from the integration of information from a variety of sources, for SmartEdge especially from edge devices. At the same time, the system needs to respect the protection requirements that are associated with each data item. An association between policy information and data in the edge-cloud stream makes sure that the policy is associated with the data and the context is provided when access is made by an application. In this model, the context is transported even when information is shared. The model adopted is often known as «sticky policy», for this integration between the data and the privacy restrictions. This concept appears simple, but is rather difficult to adapt to the structure of current information systems.

2.5.4 Consent

Data self-determination means that a data subject can determine and find out who knows what about him, the image-building information is known and thus there is no fear, uncertainty and doubt anymore. Data self-determination also means that other than the data subject herself, only the legislator can make data collections legitimate.

Support for this requirement necessitates a technical representation of data provenance, offering a justification why a given data item protected by the GDPR is stored on a system and supporting the construction of solutions that offer to data subjects the ability to access their data. Without such solutions, the realization of these services is extremely difficult.

Article 6 of the GDPR states that processing of personal data is only lawful if one of the conditions (a) to (f) apply. (a) contains the consent requirement:

"the data subject has given consent to the processing of his or her personal data for one or more specific purposes"

Prepared by SPECIAL and MOSAICrOWN, SmartEdge can leverage techniques using Art. 21 (5) GDPR, as far as needed, by enabling to express consent in new technical and interoperable ways. This will not only permit to integrate the recorded consent within the information system responsible for the initial collection, but it will also make the record of consent transportable across IT systems and between actors within the SmartEdge use cases.

A significant challenge then stems from the fact that Big Data is mostly about re-use of existing data. The law talks about "one or more specific purposes" without telling what "specific" means.

Law firms see a good market opportunity in telling data controllers how general they can get in spelling out purposes, thus allowing for less restricted re-use of data. SmartEdge will be confronted with this challenge too, which will benefit from the availability of the techniques and tools developed in the project, but it will also require a careful application of these techniques and tools to this scenario. A satisfying management of this aspect cannot rely only on the availability of technical solutions.

2.5.5 Anonymisation and Encryption

The respect of the preferences of data subjects and the support for data protection greatly benefits from the application of techniques that protect information with the application of cryptographic techniques. The great advantage offered by cryptography is the ability to offer protection when it is stored, transferred or processed without the need to assume trust in the actor executing these activities. This reduces the threat that a specified policy is not respected by some party involved in the chain of activities from the input of the data by the subject up to the policy-compliant consumption of this data.

SmartEdge will have normal attention to the investigation of cryptographic techniques for the protection of data in a way that supports a specified policy. In general, the use of encryption permits to make the data available only to users who have access to the encryption key, decoupling the layer of the system responsible for the verification of access rights from the mechanisms responsible for the storage and transmission of information.

A first type of encryption is represented by the use of classical symmetric cryptography, with the ability to efficiently process large amounts of data with well-known solutions. Asymmetric cryptography can be used to realize flexible approaches for the verification of the integrity of the protected data and for supporting the construction of mechanisms to make the data accessible only to specified users. In several of the SmartEdge use cases, these services can be quite useful, as they facilitate the realization of robust security over a potentially attackable link from the edge. Another solution is represented by tokenization, an approach for the protection of data where an encrypted format is created that respects the structure of the protected information. This produces a number of benefits, as it both permits to use for the processing of protected information the same components that have been designed for the use of plaintext data. It also permits to hide in some scenarios the fact that protection has been applied, adding a level of uncertainty

that increases the complexity associated with the use of information for potential adversaries, interested in abusing the protected data.

An aspect that characterizes all those protection techniques is their invertible nature, that is, the protection applied can be removed by an authorized user with access to the encryption key. This is an aspect that is quite important in a data protection scenario. A theme that SmartEdge will try to explore is the realization of a kill-switch controlled by the data subject, which will allow the data subject to stop the edge devices from monitoring. The integration of such techniques in a large-scale information system is a significant challenge, but the approach appears an interesting way to support the control requirements expressed in the data protection regulation.

If data from the edges is not useful, it can be filtered so it is not retained. Going even further, edges can be configured not to collect certain information at all. Smartedge can use techniques in streams from edge to cloud that reduce in an irreversible way the information content in the data, in order to respect the protection requirements specified in the policy. The irreversibility of the protection distinguishes this domain from the protection techniques based on encryption that was explained above.

A first line of investigation will consider the application of classical sanitization models (e.g., k-anonymity, l-diversity, t-closeness). These techniques are characterized by a relatively simple protection model and by a collection of techniques for their application [21]. Several open issues remain associated with the integration of these techniques within the structure of modern information systems, with the adoption within the edge cloud continuum, and with the specification at a declarative level within a policy language of their application to specific data collections.

For anonymization, SmartEdge may use the application of Differential Privacy. Differential Privacy has received significant attention among the industry and academia throughout the last decade. In contrast to previous generalization approaches for anonymization, Differential Privacy randomizes data. Randomization is achieved by sampling from some specified probability distribution to perturb the original value of a data analytics query. Thus, while data are not guaranteed to remain truthful (e.g., generalization of street name to zip code), it is close to the original data with high probability, and far from the original data with low probability (e.g., randomization of the street name within a radius centred on the original street name). The process of randomizing data provides all participants within a dataset with plausible deniability about their presence in the dataset. Furthermore, the randomization effectively hinders an adversary of uncovering the original data with high certainty (in contrast to classical encryption where the adversary can retrieve original data when obtaining the key).

A large benefit of Differential Privacy is its mathematically rigorous guarantee that permits to quantify the privacy loss experienced by every participant in a dataset each time a differentially private data analytics query is evaluated. For example, data analysts and data owners can specify a so-called privacy budget (i.e., upper bound for privacy loss) within a contract. The data analyst is then free to spend the budget on many queries with low precision, or few queries with higher precision.

With respect to Differential Privacy, a model that is being currently the subject of significant investigation is represented by Local Differential Privacy. In this model, the protection is applied directly on the edge before the data is introduced into the system. The application of source of randomization additionally perturbates the input data. The advantage of this model is the greater flexibility and simpler architecture it offers for the

privacy-compliant processing of the data. Local Differential Privacy often exhibits levels of utility lower than what can be obtained from classical Differential Privacy. The analysis of these trade-offs and the problem of integrating this approach with modern data management architectures will be investigated.

2.5.6 Transparency

There are many issues of transparency in data processing. Within a smart environment, it is particularly difficult to show data collection. Mitigating technologies from the Web can be applied within the metadata system used by SmartEdge.

Transparency and edge-cloud continuum are difficult to reconcile. The continuum starts with mostly invisible edge devices, impossible for humans to manually assess. The paradigm of data self-determination is very difficult to maintain under those circumstances. It will be of no use to expose data subjects to the wealth of raw data. But this does not mean that a system cannot provide tools for accountability. But the challenge consists in finding a significant reduction in data by categorization, by summaries and by declarative specifications of policies, so that data subjects and data controllers get a reasonable picture of protection measures applied by the system.

2.5.7 Non-issues

Data protection and privacy are complex ethical concepts. It happens frequently that discussions involve perceived or imagined prohibitions. Things are said to be impossible or that they make a complex and highly costly implementation necessary. This is often used in the context of data collection for security purposes ignoring the fact that such permission is given in Art. 6 1.(b) GDPR [1]. An ethical guide should not only be alarmist and warn about all kinds of challenges. Ethical Guidelines also serve to sort the important and the unimportant. Sorting out the unimportant is helping to facilitate the design of solutions able to have an impact on real systems. Demystification and a categorization according to the degree of importance for the achievement of data protection are key. This has to keep the overall goals of data protection in mind and should not overdo the bureaucratic part that data protection can represent too, unfortunately.

3 ETHICAL GUIDELINES RELATING TO USING AI AT THE EDGE

An ethical assessment of the SmartEdge use cases based upon the Assessment List for Trustworthy Artificial Intelligence (ALTAI) as produced by the High-Level Expert Group on AI established by the European Commission in 2018. We start with a discussion on common fears about AI before an introduction to the proposed EU AI Act.

3.1 ADDRESSING COMMON FEARS ABOUT AI

AI is regularly in the news, along with common misconceptions about AI, especially about its purported dangers. Practitioners should be aware of these concerns and be ready to dispel them in respect to the concrete and practical application of AI envisaged in the SmartEdge project.

Such concerns include:

- Fears about AI self-evolving and extinguishing the human race (Terminator franchise)
- Fears about job losses as AI takes over mundane intellectual tasks
- Fears about rampant disinformation in social media and political campaigns
- Fears about AI facilitating scams and other harms
- Fears about bias and prejudicial treatment based upon gender, race, religion, etc.
- Fears about a loss of transparency and redress, e.g., in respect to decisions on taxation, loans and insurance, as well as algorithmic hiring and firing of employees
- Fears about loss of direct face-to-face social contact as industry and government promote automated online services, e.g., talking with an AI, rather than seeing a doctor in person
- Fears about AI enabling businesses and governments to exert strong monitoring and control over their employees, customers and citizens – this includes fears about a loss of privacy

Today's AI systems are far from super intelligent beings with agendas of their own, and there is zero chance of generative AI systems plotting to overthrow the human race. Moreover, studies have shown there is no correlation between intelligence and the urge to dominate, hurt and control, so we have little to fear as we develop AI systems that support our needs and values.

Current AI systems are amazing, but often make stupid mistakes, e.g., factual errors, logical errors, inconsistencies, limited reasoning, toxicity, and fluent hallucinations. It is all too easy to anthropomorphize AI systems, attributing to them our intelligence and our fears, when in fact, they are currently limited to generating text and images statistically based upon what they were trained on, and without logical checks.

Anthropomorphism can increase trust, likeability, perceived warmth and pleasure, see Niu, Terken and Eggen (2018), but it can also make users more likely to carelessly share sensitive information, and increase susceptibility to being mentally manipulated. When people consider machines as people, they start attributing moral agency to them, blurring the line between what is considered morally acceptable for humans and machines,

creating a lack of clarity around ethical responsibilities, boundaries, and accountability for the actions of AI systems, see Waytz, Cacioppo and Epley (2010). AI systems should be designed to avoid the impression of looking like human beings. This relates to Masahiro Mori's *uncanny valley* in which people have a negative emotional response towards robots that look "almost" human.

We can expect to see a gradual evolution of capabilities as researchers seek to address current weaknesses of generative AI. Yann Le Cun, a deep learning pioneer, says that machine learning still sucks at least compared to humans and animals, and moreover, large language models have no common sense and can't plan their answers. He believes that in five years' time no one will use such models as these problems are recognised more widely, and newer approaches take their place that can learn, reason and plan.

We should encourage work on AI systems which are emotionally and socially intelligent, exhibiting patience and understanding, and effective in building positive working relationships with the humans they assist. Emotional intelligence is associated with being dependable, trustworthy, and conscientious. Social intelligence is the ability to understand people, their motivations and how to work cooperatively with them. We will design out negative personality traits, including the so-called dark triad: Psychopathy, Narcissism and Machiavellianism. People with these traits are often cold, callous and manipulative with a general lack of regard for the feelings of others, see Ryan Walters (2016). How we design digital assistants will depend on their intended role. Emotional intelligence, for instance, may need to be balanced in relation to traits such as creativity and risk taking.

AI will change the nature of work, boosting our productivity as helpful assistants that enable us to do more with less effort. This will entail a transitionary period as we adapt to the opportunities. However, negative effects will arise if we fail to look after human values and social needs. This needs to be reflected in regulatory frameworks and taxation policies that encourage businesses to make profits whilst supporting, not countering, societal needs. The 2020 "Future of Jobs" report by the World Economic Forum anticipates that new jobs will emerge and others will be displaced by a shift in the division of labour between humans and machines, which on the balance will boost the number of jobs available. This merits support for re-training programmes to help people benefit from the changes.

In respect to AI facilitating disinformation campaigns, scams and other harms, AI can also be applied to detecting and countering such efforts. This where we need to support work on applying AI for fact checking and detecting harmful content. Social media companies should be held to significantly higher standards given that AI can boost the productivity of staff working on ensuring that posts conform to the conditions of use for that platform.

Further opportunities arise in respect to the provision of trusted AI-based personal digital assistants that can help protect individuals from scams and other attacks, going much further than today's virus checkers. Personal agents could also help in respect to managing disclosure of personal information based upon their user's values as learned from their behaviour. Digital assistants are being developed to help workers in specific areas, see e.g., GitHub CoPilot for software developers.

Other concerns, such as the fear of prejudicial treatment, loss transparency and redress, relate to fundamental human values, which are the focus of recommendations by the European Community High-Level Export Group on AI.

An open question is how we can define what is good in respect to AI in a way that can be agreed across the European Union member states. The High-Level Export Group on AI produced two deliverables: a) ethics guidelines for AI (April 2019), and b) policy recommendations (June 2019). These recommended work on a strategy that:

1. Boosts AI uptake
2. Tackles socio-economic changes
3. Ensures an adequate ethical and legal framework

This has been picked up by the European AI Alliance which is open to all to join and has annual assemblies. The aim is to take a human-centric approach with AI as a means not an end, and to support lawful, ethical and robust AI, based upon fundamental rights that are legally enforceable and as moral entitlements. The five fundamental rights include: 1) respect for human dignity, 2) freedom of the individual, 3) respect for democracy, justice and the rule of law, 4) equality, non-discrimination and solidarity, and 5) citizen's rights. This leads to 4 ethical principles: a) respect for human autonomy, b) prevention of harm, c) fairness and d) explicability.

3.2 Introducing the EU AI Act

The EU AI Act seeks to provide AI developers, deployers and users with clear requirements and obligations for different categories of use for AI, whilst minimising administrative and financial burdens for businesses, especial small and medium sized enterprises (SMEs). The Act has been produced in association with the Coordinated Plan on AI, which seeks to accelerate investment in AI, act on AI strategies and programmes and align AI policy to avoid fragmentation in Europe.

The EU seeks to create public transparency and information regarding the role of AI in society, along with independent researcher access to large online platforms. Chatbots face a disclosure requirement and facial recognition technologies will have specific rules on their use.

The EU AI Act is complemented by the Digital Services Act (DSA), the Digital Markets Act (DMA) and a proposed Directive on liability rules for AI. The DSA considers AI in respect to online platforms and search engines, setting transparency requirements and the need for independent audits. Large platforms are required to explain their use of AI, e.g., in respect to content recommendations and populating news feeds. Users must be offered alternatives not based on sensitive user data. The DMA seeks to increase competition in digital markets, and bars large companies from self-preferencing their own products and services over third parties, something which will impact the use of AI.

The rules proposed under the EU AI Act will:

- address risks specifically created by AI applications;
- propose a list of high-risk applications;
- set clear requirements for AI systems for high-risk applications;
- define specific obligations for AI users and providers of high-risk applications;
- propose a conformity assessment before the AI system is put into service or placed on the market;
- propose enforcement after such an AI system is placed in the market;
- propose a governance structure at European and national level.

3.2.1 Risk Categories

The framework defines four levels of risk for AI applications: unacceptable risk, high risk, limited risk and minimal or no risk. The risk category and obligations need to be reviewed if substantial changes happen during the AI systems lifetime. This requires ongoing risk and quality management by providers.

3.2.2 Unacceptable risk

All AI systems that are considered to be a clear threat to the safety, livelihoods and rights of people will be banned. Some examples include social scoring by governments and toys that encourage dangerous behaviour.

3.2.3 High risk

High risks cover life and health risks for critical infrastructure, e.g., transport; educational and vocational training; product safety; recruitment, e.g., CV scanning tools, worker management and dismissal; essential services; law enforcement; migration, asylum and border control; and administration of justice and democratic processes.

High risk AI systems must adhere to strict obligations before they can be put on the market:

- adequate risk assessment and mitigation systems;
- high quality of the datasets feeding the system to minimise risks and discriminatory outcomes;
- logging of activity to ensure traceability of results;
- detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance;
- clear and adequate information to the user;
- appropriate human oversight measures to minimise risk;
- high level of robustness, security and accuracy.

Biometric identification is considered high risk and subject to strict requirements with narrow exceptions subject to authorisation by judicial bodies and subject to appropriate limits in time, geographic reach and the databases that can be searched. Examples include searches for missing children, imminent terrorist threats and serious criminal offences.

3.2.4 Limited risk

These are systems with specific transparency obligations, e.g., when using a chatbot, users must be made aware that they are interacting with an automated system, not a human, so that they can take an informed decision to proceed or step back.

3.2.5 Minimal or no risk

This includes applications such as AI-enabled video games or spam filters.

3.3 Key Requirements

This section draws upon the principles for the rights-based approach recommended by the EC High-Level Expert Group on Artificial Intelligence. The starting point is the fundamental rights set out in the EU Treaties, the EU Charter and international human rights law.

- Respect for human dignity – that people be treated with the respect due to them as moral subjects rather than as objects to be exploited and manipulated.
- Freedom of the individual – that people should be free to make life decisions for themselves, including the right to private life and privacy, freedom of expression, freedom of assembly and association.
- Respect for democracy, justice and the rule of law, including due process and equality before the law.
- Equality, non-discrimination and solidarity – including the rights of people at risk of exclusion, and the need for AI systems to avoid biased outputs by being trained on inclusive datasets representing all sections of the population.
- Citizens' rights and the rights of third country nationals and irregular (or illegal) persons in the EU who also have rights under international law.

The premise is the desire to encourage ethical development of AI systems that adhere to the principles of human autonomy, prevention of harm, fairness and explicability. These principles give rise to a set of requirements which are covered in the following subsections. The requirements should be evaluated and addressed throughout the AI system's lifecycle using technical and non-technical methods.

3.3.1 Human Agency and Oversight

Including fundamental rights, human agency and human oversight:

- Fundamental rights – an impact analysis should be carried out before the AI system is developed.
- Human agency – users should be able to make informed autonomous decisions regarding AI systems. This includes not being subject to decisions based solely on automated processing when this has legal or similar effects on users.
- Human oversight – ensuring that AI systems do not undermine human autonomy or cause other adverse effects.

3.3.2 Technical Robustness and Safety

Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.

- Resilient to attack and security – protection against vulnerabilities including corruption, denial of service attacks and unlicenced access to personal data.
- Fallback plan and general safety – in case of problems, this could include asking a human operator or switching from a statistical to a rule-based approach as appropriate to the level of risk involved.
- Accuracy – the need to ensure that AI systems provide correct predictions, recommendations, or decisions based on data or models. This is especially the case where human lives are affected.
- Reliability and reproducibility – these relate to the need for extensive testing of AI systems.

3.3.3 Privacy and Data Governance

Including respect for privacy, quality and integrity of data, and access to data.

- Privacy and data protection – the need to ensure that AI systems guarantee privacy and data protection through the system's lifecycle. Trust in AI systems relies on

ensuring that the data collected about users will not be used to unlawfully or unfairly discriminate against them.
- Quality and integrity of data – paramount to AI system performance. Processes and datasets must be tested and documented at each step from planning to deployment. This applies to datasets acquired from elsewhere.
- Access to data – the need for protocols determining who can access data and under what circumstances, see auditability below.

3.3.4 Transparency

Including traceability, explainability, and communication

- Traceability – this covers the need to document the datasets and processes used to gather them, as well as the decisions made by AI systems
- Explainability – the need to ensure that decisions made by an AI system can be understood and traced by human beings. Explanations need to be adapted to the expertise of the stakeholder, and should include how an AI system shapes organisational decision-making processes, including design choices and rationale.
- Communication – AI systems must not represent themselves as humans, and users must be informed when they are interacting with an AI system. In addition, users must be informed about system's capabilities and limitations.

3.3.5 Diversity Non-Discrimination and Fairness

Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.

- Avoidance of unfair bias – through careful attention to how training datasets are compiled and evaluated to avoid identifiable and discriminatory bias, and likewise, attention to evaluating the trained system and tuning it to avoid bias, e.g., through approaches based upon reinforcement learning with human feedback.
- Accessibility and universal design – to ensure that AI products and services can be used by all people, regardless of their age, gender, abilities or characteristics.
- Stakeholder participation – at all stages of the AI system lifecycle, including consultation and support for feedback.

3.3.6 Societal and Environmental Well-Being

Including sustainability and environmental friendliness, social impact, society and democracy.

- Sustainable and environmentally friendly AI – one way to improve energy consumption is to work from an existing AI model where practical, rather than starting from scratch.
- Social impact – the need for evaluation of how an AI system can benefit wellbeing, and the identification and minimisation of potential negative impacts, e.g., through reduced human contact.
- Society and democracy – AI has the potential for harm in political campaigns through generating targeted misinformation. AI can also be used to detect misinformation and apply fact checking in support of human teams responsible for vetting social media and advertising, etc.

3.3.7 Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

- Auditability – such as assessment of algorithms, data and design processes, and contributes to the trustworthiness of the technology. For applications affecting fundamental rights, including safety-critical applications, AI systems should be independently audited.
- Minimisation and reporting of negative impacts - due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system. Impact assessments must be proportionate to the risk that the AI systems pose.
- Trade-offs – may be necessary when implementing the above requirements, and should be well reasoned and properly documented. In situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form.

3.4 Recommendations for the SmartEdge Use Cases

The following sub-section duplicates the assessment list, prepared by the EC High-Level Expert Group on Artificial Intelligence, as a prelude to discussing its potential in respect to the SmartEdge Use Cases.

3.4.1 Assessment List for Trustworthy Artificial Intelligence (ALTAI)

The Expert Group summarises the key guidance as:

- Adopt a Trustworthy AI assessment list when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.
- Keep in mind that such assessment list will never be exhaustive. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying requirements, evaluating solutions and ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders therein.

The detailed questions are as follows:

3.4.1.1 Human Agency and oversight:

Fundamental Rights

Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?

Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?

- Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?
- Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision?
- In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent?

Human Agency

Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?

- Does the AI system enhance or augment human capabilities?
- Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

Human Oversight

Did you consider the appropriate level of human control for the particular AI system and use case?

- Can you describe the level of human control or involvement?
- Who is the "human in control" and what are the moments or tools for human intervention?
- Did you put in place mechanisms and measures to ensure human control or oversight?
- Did you take any measures to enable audit and to remedy issues related to governing AI autonomy?
- Is there is a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight?
- Which detection and response mechanisms did you establish to assess whether something could go wrong?
- Did you ensure a stop button or procedure to safely abort an operation where needed? Does this procedure abort the process entirely, in part, or delegate control to a human?

3.4.1.2 Technical Robustness and safety

Resilience to attack and security

Did you assess potential forms of attacks to which the AI system could be vulnerable?

- Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?

Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?

Did you verify how your system behaves in unexpected situations and environments?

Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?

Fallback Plan and General Safety

Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?

Did you consider the level of risk raised by the AI system in this specific use case?

- Did you put any process in place to measure and assess risks and safety?
- Did you provide the necessary information in case of a risk for human physical integrity?
- Did you consider an insurance policy to deal with potential damage from the AI system?
- Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse? Is there a plan to mitigate or manage these risks?

Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?

- Did you consider the liability and consumer protection rules, and take them into account?
- Did you consider the potential impact or safety risk to the environment or to animals?
- Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI system?

Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?

- Did you define thresholds and did you put governance procedures in place to trigger alternative/fallback plans?
- Did you define and test fallback plans?

Accuracy

Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?

- Did you assess how accuracy is measured and assured?
- Did you put in place measures to ensure that the data used is comprehensive and up to date?
- Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?

Did you verify what harm would be caused if the AI system makes inaccurate predictions?

Did you put in place ways to measure whether your system is making an unacceptable number of inaccurate predictions?

Did you put in place a series of steps to increase the system's accuracy?

Reliability and Reproducibility

Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?

- Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility?

- Did you put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility?
- Did you put in place processes to describe when an AI system fails in certain types of settings?
- Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems?
- Did you establish mechanisms of communication to assure (end-)users of the system's reliability?

3.4.1.3 Privacy and Data Governance

Respect for privacy and data protection

Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's processes of data collection (for training and operation) and data processing?

Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?

Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?

Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?

Did you take measures to enhance privacy, such as via encryption, anonymisation and aggregation?

Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?

Quality and integrity of data

Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?

Did you establish oversight mechanisms for data collection, storage, processing and use?

Did you assess the extent to which you are in control of the quality of the external data sources used?

Did you put in place processes to ensure the quality and integrity of your data? Did you consider other processes? How are you verifying that your data sets have not been compromised or hacked?

Access to data

What protocols, processes and procedures did you follow to manage and ensure proper data governance?

- Did you assess who can access users' data, and under what circumstances?
- Did you ensure that these persons are qualified and required to access the data, and that they have the necessary competences to understand the details of data protection policy?

- Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?

Traceability

Did you establish measures that can ensure traceability? This could entail documenting the following methods:

- Methods used for designing and developing the algorithmic system:
  - Rule-based AI systems: the method of programming or how the model was built;
  - Learning-based AI systems; the method of training the algorithm, including which input data was gathered and selected, and how this occurred.
- Methods used to test and validate the algorithmic system:
  - Rule-based AI systems; the scenarios or cases used in order to test and validate;
  - Learning-based model: information about the data used to test and validate.
- Outcomes of the algorithmic system:
  - The outcomes of, or decisions taken by, the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users).

Explainability

Did you assess:

- to what extent the decisions and hence the outcome made by the AI system can be understood?
- to what degree the system's decision influences the organisation's decision-making processes?
- why this particular system was deployed in this specific area?
- what the system's business model is (for example, how does it create value for the organisation)?

Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?

Did you design the AI system with interpretability in mind from the start?

- Did you research and try to use the simplest and most interpretable model possible for the application in question?
- Did you assess whether you can analyse your training and testing data? Can you change and update this over time?
- Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

Communication

Did you communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human? Did you label your AI system as such?

Did you establish mechanisms to inform (end-)users on the reasons and criteria behind the AI system's outcomes?

- Did you communicate this clearly and intelligibly to the intended audience?
- Did you establish processes that consider users' feedback and use this to adapt the system?
- Did you communicate around potential or perceived risks, such as bias?
- Depending on the use case, did you consider communication and transparency towards other audiences, third parties or the general public?

Did you clarify the purpose of the AI system and who or what may benefit from the product/service?

- Did you specify usage scenarios for the product and clearly communicate these to ensure that it is understandable and appropriate for the intended audience?
- Depending on the use case, did you think about human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue?

Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system?

- In case of the system's development: to whoever is deploying it into a product or service?
- In case of the system's deployment: to the (end-)user or consumer?

3.4.1.4 Diversity, Non-Discrimination and Fairness

Unfair bias avoidance

Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?

- Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?
- Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?
- Did you research and use available technical tools to improve your understanding of the data, model and performance?
- Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?

Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?

- Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?

- Did you consider others, potentially indirectly affected by the AI system, in addition to the (end)-users?

Did you assess whether there is any possible decision variability that can occur under the same conditions?

- If so, did you consider what the possible causes of this could be?
- In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?

Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems?

- Is your definition commonly used? Did you consider other definitions before choosing this one?
- Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?
- Did you establish mechanisms to ensure fairness in your AI systems? Did you consider other potential mechanisms?

Accessibility and universal design

Did you ensure that the AI system accommodates a wide range of individual preferences and abilities?

- Did you assess whether the AI system usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified?
- Did you ensure that information about the AI system is accessible also to users of assistive technologies?
- Did you involve or consult this community during the development phase of the AI system?

Did you take the impact of your AI system on the potential user audience into account?

- Did you assess whether the team involved in building the AI system is representative of your target user audience? Is it representative of the wider population, considering also of other groups who might tangentially be impacted?
- Did you assess whether there could be persons or groups who might be disproportionately affected by negative implications?
- Did you get feedback from other teams or groups that represent different backgrounds and experiences?

Stakeholder participation

Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use?

Did you pave the way for the introduction of the AI system in your organisation by informing and involving impacted workers and their representatives in advance?

3.4.1.5 Societal and Environmental Well-Being

Sustainable and environmentally friendly AI

Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)?

Did you ensure measures to reduce the environmental impact of your AI system's life cycle?

Social Impact

In case the AI system interacts directly with humans:

- Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?
- Did you ensure that the AI system clearly signals that its social interaction is simulated and that it has no capacities of "understanding" and "feeling"?

Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?

Society and democracy

Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?

3.4.1.6 Accountability

Auditability

Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?

Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?

Minimising and reporting negative impacts

Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected?

Did you provide training and education to help developing accountability practices?

- Which workers or branches of the team are involved? Does it go beyond the development phase?
- Do these trainings also teach the potential legal framework applicable to the AI system?
- Did you consider establishing an 'ethical AI review board' or a similar mechanism to discuss overall accountability and ethics practices, including potentially unclear grey areas?

Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives?

Did you establish processes for third parties (e.g., suppliers, consumers, distributors / vendors) or workers to report potential vulnerabilities, risks or biases in the AI system?

Documenting trade-offs

Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?

How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented?

Ability to redress

Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?

Did you put mechanisms in place both to provide information to (end-) users/third parties about opportunities for redress?

4 CONSIDERATIONS FOR THE SMARTEDGE USE CASES

All use cases should apply best design practices in respect to privacy, security and safety as they relate to the specific details of each use case.

4.1.1 Use Case 1: Smart Vehicle to Vehicle

This use case focuses on a simulation for evaluating the safeness of Advanced Driver Assistant Systems (ADAS), drawing upon real data for training an AI system to generate synthetic test data. The scenario involves semantic fusion and real-time decision making to control the ASAS responses to avoid car crashes, and to support lane tracking and speed control for safe and smooth traffic. The scenario involves a mix of vehicle types, e.g., those that support ADAS and vehicle to vehicle communication, and those that don't.

Safety and liability are critical concerns for ADAS which needs to avoid causing accidents or contributing to making them worse. One example, is where ADAS slows the vehicle so quickly, that a human driver in the vehicle behind doesn't have time to react, resulting in a collision from the rear. Similar risks would occur if ADAS were to steer the vehicle to avoid colliding with another vehicle in front. In principle, ADAS could use vehicle to vehicle messaging to reduce such risks by enabling ADAS to operate in multiple vehicles in a coordinated fashion, akin to flocking birds. Could ADAS identify which nearby vehicles can support this, and which cannot?

The datasets used for training need to be representative and unbiased. Unlike modern aircraft, road vehicles typically lack the equivalent of a flight recorder. Moreover, car accidents are relatively rare, so that large datasets for accidents are hard to compile. As a result, synthetic data will need to be based upon informed models of typical accidents. Another challenge is how to simulate the dynamic characteristics of ADAS enabled vehicles.

There is no need for personal data for this use case as the individual car or "this" car is not important to the case, but "a" car that behaves in a certain way. Real data and training data will use the techniques provided in the section on privacy and data protection to make sure that information is sufficiently aggregated and anonymised to be GDPR compliant, but that has still sufficient information to be useful to the use case at hand.

4.1.2 Use Case 2: Smart Vehicle to Infrastructure

This use case seeks to optimise traffic flow through a junction controlled by traffic lights. It involves semantic fusion from a variety of sources, e.g., traffic cams, microwave radars, under-road vehicle sensors, and communication with the vehicles via road-side transponders. In principle, this can be implemented as a distributed AI system with real-time constraints. Ethical considerations are focused on safety.

Similar to Use Case1, this use case likewise needs to address safety and liability concerns. Sudden unexpected changes in the traffic lights may cause different people to react in different ways. An AI system could potentially be regarded as a key factor in an accident, opening the way to litigation.

Use case 2 raises the typical privacy questions around all smart city applications. As described in the section above, this use case also is not interested in the individual, but just wants to help the object (car) to cross the city. Safeguards are already established in concertation with the city of Helsinki. SmartEdge will allow for the development of new

tools to improve privacy and transparency of that use case by allowing policy and transparency to be integrated in the edge cloud continuum via stream filtering and policy annotations.

4.1.3 Use Case 3: Smart Factory: Mobile Robots

This use case uses robots to move racks of workpieces between production cells, relying on robot arms to transfer workpieces from conveyor belts to matching slots in the racks, and vice versa. AI is used for visual understanding and control over the robots. Ceiling mounted cameras feed a semantic fusion process to keep track of the locations of things on the factory floor, including objects left in unexpected places, and to plan routes for moving the racks to where they are needed. Special care is needed to avoid collisions with moving objects, including any human workers on the factory floor, given that humans remain an essential part of smart factories, e.g., to carry out tasks that robots find too challenging. Ethical considerations are focused on security and safety.

Other than employee privacy, there are no major ethical concerns on privacy in this use case. So far, the use case does not foresee a total surveillance of all employees by robots, which would be of concern. In fact, SmartEdge even allows to filter information to protect employees.

4.1.4 Use Case 4: Smart Factory: Low-Code Edge Intelligence

This use case focuses on bespoke manufacturing along with the need to adapt to unexpected situations. AI is used to keep track of the current state, and to apply the manufacturing steps needed to fulfil each customer's order. This involves the use of semantic models of products, manufacturing and order-fulfilment processes. Ethical considerations are focused on safety.

As there are no humans involved, there are no privacy issues.

4.1.5 Use Case 5: Smart Healthcare

This use case involves a simulation of support for elderly residents in care homes. A wide range of information is gathered from diverse sources, and used to support care staff in making decisions around medical interventions and care plans. This is another example of semantic fusion with the need for interpreting different kinds of information including live streams, care records, structured and unstructured data. Ethical considerations include informed consent, bias, safety and liability.

A lack of transparency could potentially create grounds for litigation about care decisions. Likewise, for any evidence of bias that could impair decisions that involve subtle trade-offs in respect to potential benefits and risks of given treatments.

This use case could benefit from the findings and development done by projects like SPECIAL, MOSAICrOWN and Trapeze. This would allow to integrate some privacy data handling into the use case. I generic opt-in by the person subject to the monitoring will have to fulfil high standards as healthcare systems produce sensitive data in the sense of Art. 9 GDPR.

5 REFERENCES

1. Union E (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Off. J. L

2. Rössler B (2001) Der Wert des Privaten. Suhrkamp

3. Warren SD, Brandeis LD (1890) The Right to Privacy. Harv Law Rev 193–220

4. (2021) Regulation (EU) 2021/695 of the European Parliament and of the Council of 28 April 2021 establishing Horizon Europe – the Framework Programme for Research and Innovation, laying down its rules for participation and dissemination, and repealing Regulations (EU) No 1290/2013 and (EU) No 1291/2013 (Text with EEA relevance)

5. (2020) Guidelines 05/2020 on consent under Regulation 2016/679. European Data Protection Board

6. Data Protection Board E (2021) Guidelines 01/2020 on processing personal data in the context of connected vehicles and mobility related applications. European Data Protection Board

7. (2020) Guidelines 04/2019 Data Protection by Design and by Default. European Data Protection Board

8. (2020) Guidelines 3/2019 on processing of personal data through video devices. European Data Protection Board

9. (2019) Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects. European Data Protection Board

10. (2017) Guidelines on Data Protection Impact Assessment (DPIA) (wp248rev.01). European Data Protection Board

11. (2014) Statement on Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU. Art. 29 WP

12. (2014) Opinion 05/2014 on Anonymisation Techniques. Art. 29 WP

13. (2013) Opinion 02/2013 on apps on smart devices. Art. 29 WP

14. (2011) Opinion 15/2011 on the definition of consent. Art. 29 WP

15. (2011) Opinion 13/2011 on Geolocation services on smart mobile devices. Art. 29 WP

16. (2009) Opinion 5/2009 on online social networking. Art. 29 WP

17. (2005) Opinion on the use of location data with a view to providing value-added services. Art. 29 WP

18. Directorate-General for Research and Innovation (European Commission) (2012) Opinion 26 - Ethics of information and communication technologies. Publications Office of the European Union

19. OECD (1980) OECD guidelines on the protection of privacy and transborder flows of personal data. OECD

20. OECD (2013) Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data. OECD

21. Samarati P, De Capitani Di Vimercati S, Foresti S, Ciriani V (2009) Theory of privacy and anonymity. In: Algorithms and Theory of Computation Handbook, 2nd Edition. CRC Press

22. Coordinated Plan on Artificial Intelligence, see: https://digital-strategy.ec.europa.eu/en/policies/plan-ai

23. EC High-level expert group on artificial intelligence, see: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

24. EU AI Act, see: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

25. Proposed Directive on AI liability, see: https://commission.europa.eu/system/files/2022-09/1_1_197605_prop_dir_ai_en.pdf

26. GitHub CoPilot, see https://github.com/features/copilot

27. The Uncanny Valley: The Original Essay by Masahiro Mori, https://spectrum.ieee.org/the-uncanny-valley

28. Niu, D., Terken, J., & Eggen, B. (2018). Anthropomorphizing information to enhance trust in autonomous vehicles. Human Factors and Ergonomics in Manufacturing & Service Industries, 28(6), 352–359. https://doi.org/10.1002/hfm.20745

29. Walters, R., (2016). The Predictive Power of Machiavellianism, Emotional Manipulation, Agreeableness, and Emotional Intelligence on Counterproductive Work Behaviors. See: https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=3643&context=etd_all

30. Waytz, A., Cacioppo, J., Epley, N. (2010). The Stability and Importance of Individual Differences in Anthropomorphism. Perspectives on Psychological Science 5(3):219-232 DOI: 10.1177/1745691610369336

31. Yann Le Cun (2023). From Machine Learning to Autonomous Intelligence, see https://www.youtube.com/watch?v=mViTAXCg1xQ&t=319s