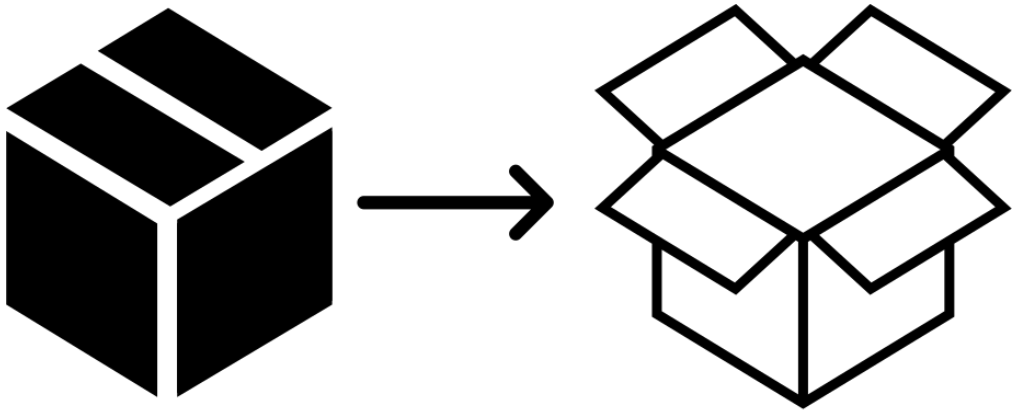
chist-era



XAIface

Measuring and Improving Explainability for AI-based Face Recognition

Ethics Guidelines for AI-based Face Recognition

Deliverable number: D3.4

Version: 1.0

EURECOM Sophia Antipolis · JOANNEUM RESEARCH DIGITAL · universität wien · it · EPFL

Acronym of the project: XAIface
Title of the project: Measuring and Improving Explainability for AI-based Face Recognition.
Grant: CHIST-ERA-19-XAI-011
Web site of the project: https://xaiface.eurecom.fr/

Short abstract
This document concerns the ethical considerations and resulting recommendations towards ethical guidelines for face recognition technology based on the work of the Artificial Intelligence High Level Expert Group (AI-HLEG) and scientific literature concerning Ethics. Ethical methodology for casuistic approaches towards use case assessment is presented next to the basic principles for the ethical application of Artificial Intelligence as proposed by the AI-HLEG. This document also discusses potential loopholes in current laws and differences in existing recommendations and legal requirements, as well as their ethical resolving.

Table of content

Definitions 3

1. Introduction 4
   1.1. Face Recognition 4
   1.2. Ethics as discipline 4
      1.2.1. Definition of Terms 4
      1.2.2. Towards applied ethics in practice 5
   1.3. Disclaimer 5

2. AI-HLEG 2019 Ethics Guidelines for Trustworthy AI 6
   2.1. Principles 6
   2.2. Requirements 7
   2.3. Assessment List 8
   2.4. Face Recognition Technology - Relevant Issues 9
      2.4.1. General Considerations 9
      2.4.2. Use-case specific considerations 9

3. Preliminary Summary 10

Definitions

Biometric Data means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data.[1]

Controller is a person, who alone or jointly with others decides on the means and purposes of the data processing and can be seen as the main addressee of the GDPR.

Data subjects are natural persons, whose data will be processed.

Data Protection Directive 1995 was repealed through the GDPR and was in force until the entry of the GDPR on 25th of May in 2018.

General Data Protection Regulation (GDPR): The General Data Protection Regulation is a European legal act, which lays down European-wide harmonized provisions regarding the processing of personal data and is directly applicable in all EU-member states.[2]

Personal data is any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.[3]

Processing means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.[4]

Processor is a person who processes personal data on behalf of the controller.

[1] Art. 4(1)(14) GDPR.
[2] Regulation 2016/679/EU 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).
[3] Art. 4(1)(1) GDPR.
[4] Art. 4(1)(2) GDPR.

1. Introduction

Whenever discussing AI technology and its inherent benefits, drawbacks, and compromises, questions concerning ethics arise. Moreover, face recognition technologies, such as face validation or identification receive additional scrutiny over the societal and psychological implications of ubiquitous automated identification and its inherent potential for tracking and policing. It is thus particularly imperative that discussion about the ethical application of face recognition technologies is led in great detail when developing novel use cases and application scenarios.

1.1. Face Recognition

Face recognition technology serves the purpose of identifying persons from images. There are two general approaches: face validation, where a face is compared against a reference and the output of the algorithm is a match or mismatch, and face identification, where a person is to be identified from an image compared against a set of references. The process usually involves face detection (as in detecting the presence and location of a face on in an image), followed by feature extraction and the identification process. The technology has in recent years seen great improvement through the application of concepts machine learning or artificial intelligence and has been integrated in many products around us, be it smart phone cameras and operating systems (e.g., for security purposes), photo management apps, or surveillance camera systems.

1.2. Ethics as discipline

In layman's terms, ethics can be described as the discussion about what the "right" thing to do is. In a more scientific sense, ethics is the scientific discussion of morals, terms elaborated upon further below. Many problems arise from the fundamental questions of what is "evil" and "good", such as modern-day dilemmas derived from the infamous so-called trolley problem, and projected on modern day scenarios of self-driving cars and artificial intelligence or robots.

1.2.1. Definition of Terms

Morals are societal norms of behavior regarding social life. They are often implicit, part of culture, and can be heterogeneous even within a society. There are no perfect universal morals and differences in moral understanding can lead to conflicts. Morals can take shape as commandments or prohibitions, habits, and many more.

Ethics is the scientific philosophical examination of morals. There are many different approaches to ethics, such as descriptive ethics (non-normative scientific description of morals), or normative ethics, which are represented mostly by deontological ethics (rule-based), or utilitarianism (outcome-oriented) in Europe. Normative ethics means making value judgements and ethically arguing their validity either inductively from a set of core values, or based on the (expected) outcome of actions according to the judgement.

Ethos is a codification of morals, often created by following an ethic arguing process. Such moral codices may apply to a small field of ethics only (such as the Hippocratic Oath to medicine, or the XAIface ethical guidelines to explainable AI-based face recognition technology) others may apply in a more general sense.

1.2.2. Towards applied ethics in practice

While it would be convenient to simply check off compliance with guidelines or requirements, and thus approaching ethical questions from a top-down perspective, the heterogeneity of the possible application scenarios of face recognition technologies based on artificial intelligence mandates ethical analysis beyond checklists. This means that stakeholder involvement and integration of diverse teams into the development and assessment process throughout the life cycle of the AI system is key to truly ethical solutions. This poses the challenge of continuous specialized education and training in ethical theories, values and arguing upon them, which should ideally be available to all developers, deployers, and decision makers involved in AI and face recognition systems. When applying such a bottom-up approach, as many persons involved as possible should be able to raise concerns and ethically argue them during each step of the lifecycle of such a system in order to detect and resolve issues which otherwise pass under the radar. In practice, a combination of those approaches may be applied, with an assessment or checklist as a top-down tool for bottom up arguing.

1.3. Disclaimer

As with all ethical guidelines and considerations, this document must be viewed as non-exhaustive. While attempting to keep it as general as possible, it is impossible to anticipate all future conflicts of interest and potential imbalances in power as well as changes in a societies' morals, and thus its resulting ethos. The presented ethical guidelines are merely to be seen as tools for self-empowered ethical arguing concerning applications of face recognition technologies based on the current state of the art and values as established in the European Union.

2. AI-HLEG 2019 Ethics Guidelines for Trustworthy AI

In 2019, the European Commission's High-Level Expert Group (AI-HLEG) on Artificial Intelligence published ethics guidelines for trustworthy AI. The document proposes three pillars on which the ethic evaluation of a system throughout its lifecycle is to stand. The first pillar is the human-centric principles, which are founded on the fundamental rights, the second are key requirements for ethical AI, and the third pillar is an assessment list for the evaluation of AI-based systems and their deployments throughout their lifecycle.

2.1. Principles

The AI-HLEG Ethics Guidelines are calling for compliance with the following four principles based on fundamental rights, which also apply to AI-based face recognition technology and can thus be applied without restraint.

1. **Respect for Human Autonomy:** It is mandatory for AI systems to leave any interacting humans self-determination untouched. Unjustified influence on humans, whether deceptive, coercive, manipulative, or through subordination, conditioning, or herding should not take place. The principle of respect for human autonomy instead calls for complementation and augmentation of human capabilities concerning cognitive, social, or cultural skills.
2. **Prevention of Harm:** this principle specifies that AI systems must never exacerbate harm or create any other adverse effect on human beings. Human dignity is protected in the same way as physical integrity by this principle. In Practice, this means that AI systems must be safe and secure, as well as technically robust and hardened against abuse.
3. **Fairness:** regarding AI systems, fairness has multiple dimensions. The principle as brought forward by the AI-HLEG includes not only equal and just distribution of cost, but also ensuring freedom from unfair bias, discrimination, and stigmatization. It is noted that unbiased, fair AI systems could lead to increased societal fairness in general.
4. **Explicability:** In order to gain user trust, explicability through transparency about processes concerning as well as capabilities and purpose of AI systems is a major factor. This includes explanations as to why a specific decision was made (to the degree possible given the corresponding AI system), in order for the decisions to be contestable. If such a level of transparency cannot be achieved, other approaches such as traceability and auditability may serve as substitutions beneath open and transparent communication about the capabilities of the system.

Naturally, the aforementioned principles may come into conflict with each other. In such situations the AI-HLEG recommends approaching the ethical dilemmas with reasoned and evidence-based reflection and documenting the resulting trade-offs. Some principles, such as fundamental rights may be considered as non-negotiable though and must thus never be balanced against other principles, e.g. human dignity.

2.2. Requirements

Derived from the four principles, seven requirements for trustworthy AI were postulated by the AI-HLEG, which have to be met to the greatest possible extent in order for an AI system to be considered ethical to current European standards. All requirements need to be continuously evaluated throughout the life cycle of the AI system and can not just be checked beforehand.

1. **Human agency and oversight:** Principle of fundamental rights, human agency and human oversight. This requirement includes a fundamental rights assessment, and there need to be mechanisms for challenging the system's decision. The principle of user autonomy must be central to the functionality of the system. In order to ensure no undermining of human autonomy takes place, oversight mechanisms such as human-in-the-loop, human-on-the-loop or human-in-command need to be placed.
2. **Technical robustness and safety:** Principle of prevention of harm. This requirement mandates a preventative approach to risks, including unintentional and unexpected as well as unacceptable harm. Ways of achieving compliance with this requirement include building resilience to attack, whether that be attacks on data or the model. Dual-use, the unintended potential abuse of the system for additional applications, should be taken into account and -if possible- prevented and mitigated. There need to be fallback plans in case of problems (either rule-based or human-operation) in order to ensure the system does not harm living beings or the environment. Depending on the criticality and impact on human lives, high accuracy levels are crucial for AI systems and likeliness of errors should be indicated. AI systems should be reliable and output reproducible, meaning that they should work in a range of situations and that the same input parameters should always lead to the same outputs.
3. **Privacy and data governance:** Principle of prevention of harm. Privacy is a fundamental right, and AI systems must thus guarantee privacy and data protection throughout their lifecycle. Personal data may be initially provided to the system, such as face image datasets in the case of face recognition technology, or generated upon or during use. It must be ensured that derived sensitive data will not be used to unlawfully or unfairly discriminate against the users. The quality and integrity of the data sets strongly affects the performance of AI systems, it is thus mandatory to ensure high quality training data free from socially constructed biases, inaccuracies, errors and mistakes. The integrity of the data, and its freedom from maliciously planted entries must also be ensured. An Individual’s data should only be accessed by duly qualified personnel with the according competence and based on a specific need.
4. **Transparency:** Principle of explicability. Transparency encompasses the data, the system, and the business models. Systems need to be traceable in such a way that reasons for incorrect decisions can be identified to prevent future mistakes. This may be achieved through documentation of the important phases of data gathering and labelling, as well as the algorithms used. Explainability of both the processes within the systems, as well as the human decisions about deployment should be provided. Explanations should be made in such a way, that the involved stakeholders may understand them. Users need to be informed about the fact that they are interacting with AI systems, and options to decide against using an AI system in favor of human interactions should be provided.

5. **Diversity, non-discrimination and fairness:** Principle of fairness. Unfair bias should be avoided as it can lead to unintended prejudice and discrimination. These biases may stem from data sets with historic bias, incompleteness, or bad governance models. This requirement is to be achieved by removing the bias in the collection phase, if possible. Additionally, systems should follow accessibility design principles and be user-centric. It is important to note that stakeholders who may be affected by the system should be involved and consulted and mechanisms for long-term participation should exist. Workers in AI systems should be informed and consulted, as well as empowered to participate throughout the process.
6. **Societal and environmental well-being:** Principle of fairness and prevention of harm. The group of stakeholders regarding an AI system should be extended to society as a whole, and other sentient beings as well as the environment. A core belief stated in the recommendations is that “AI systems should be used to benefit all human beings, including future generations”. Training phases should be optimized towards resource usage and energy consumption whenever possible, and the environmental impact of supply chains may also be secured by suitable measures. Effects on the conception of social agency and social relationships need to be carefully monitored and considered. Impacts of an AI system on society and its institutions as well as democracy should be taken into account, and there need to be particular caution with regards to the democratic process in situations such as electoral processes or political decision making.
7. **Accountability:** Principle of fairness. The requirement necessitates mechanisms for responsibility and accountability in order for the compliance to previous requirements to be investigated and verified throughout all phases of the systems lifecycle. Auditability (not necessarily public availability) of algorithms, data and design processes should be given, and regular audit reports can increase trust in the technology. Negative impacts should be minimized, and reported if detected. This also includes protection for whistleblowers, NGOs, or trade unions. Trade-offs need to be documented and addressed in a rational and methodological manner implying the explicit acknowledgement of said trade-offs. In case of unjust adverse impacts, accessible mechanisms to ensure redress should be put in place in order to increase trust through correction of things that went wrong.

These seven requirements may be achieved through technical or non-technical means. The assessment list covered in the following section helps practitioners check their own plans against them.

2.3. Assessment List

Based on the requirements derived from the principles, the AI-HLEG has also published an assessment list for trustworthy AI. This assessment list is supposed to serve as a checklist for evaluating a planned AI-based system and its deployment. It is partitioned into sectors according to the requirements stated above and to be used on all levels of an organization throughout the whole lifecycle of an AI system. It is also noted that while the assessment list may go beyond what international or national law requires, it is not a sufficient check for lawfulness of an AI system. The questions posed in the assessment list are, while plentiful,

often more abstract and not easy to answer without broader discussion of the topics and in-depth analyses of potential consequences of applying the system. The assessment list is available as a PDF document and as an interactive digital tool upon registration.

2.4. Face Recognition Technology - Relevant Issues

The AI-HLEG Ethics Guidelines also raise some issues, which are particularly applicable to face recognition technology, and shall thus be discussed in this section. These Questions range from general considerations regarding multiple possible application scenarios to more specific use cases and their implications.

2.4.1. General Considerations

Inherent to all face recognition technology are the **psychological and sociocultural impacts of automatic identification.** How will the knowledge about automated identification, which is often more complicated to contest than human-made decisions, influence users on a psychological level and how will this influence manifest in our social behavior and culture? Developers and deployers of such technology will have to pay special attention to this question as each novel wide-spread application may change the perception of face recognition technology in general and thus influence human behavior when voluntarily or involuntarily interacting with it. On the topic of voluntary nature, the question of
consent or clear warrant comes into play. Since face recognition is a task performed on sensitive personal (biometric) data, the GDPR (as a codification of morals as law) already requires operators of face recognition technologies to pay special attention to either obtain consent or have another justification for the processing of the personal data covered by the GDPR. Since face recognition technologies may be deployed in public spaces, clear consent from all subjects is likely impossible to obtain, and opt-out mechanisms are unavailable. It must also be noted that an automated face recognition system may inherently **affect human decision making.** It is therefore important to keep transparency about automated decisions available to users at all times so they can make informed decisions based on that knowledge as opposed to as if they were assuming a human operator behind the automated decisions. When **Augmentation of human capabilities** takes place, such as in human-algorithm teaming for face recognition tasks, special attention has to be given to the problems of overreliance on automated decisions and overconfidence in the AI system. Good accuracy in a system may lead a human operator to reduce their effort and overly rely on the system's decisions instead of thoroughly verifying the outcomes, which can lead to additional ethical implications concerning safety and security depending on the envisioned scenario.

2.4.2. Use-case specific considerations

Other questions arise depending on specific use cases, such as using face recognition technology as a means of **identification** at one location, or using it as a means of **tracking** persons across multiple locations, which would easily be possible. On this notion, the topic of **targeted surveillance**, where a single person is to be identified, located, and surveilled, and **mass surveillance**, where this affects a large number of people also arises. If deploying such systems, measures against dual-use and abuse must be taken in order to not lose public trust in the technology.

Special attention needs to be given to **electoral contexts**, where the democratic right to vote may be affected. Scenarios of voter identification at the polls need particular consideration, since false positive or false negative results on identification may deny an eligible voter casting their vote if their identity has falsely been registered before or the identification provides no correct result.

3. Preliminary Summary

This first version of the deliverable on ethics guidelines for face recognition technology based on artificial intelligence provides a comprehensive overview over basic knowledge and terms with regards to ethics and the most important reference for ethical AI in the European union. Moreover, specific problem cases for face recognition technology have been identified and ethical questions of relevance to face recognition are posed as well as legal loopholes existing in current law. Version 2 may include casuistic approaches on concrete use cases developed in the project depending on timely availability, and specific requirements for face recognition derived from the AI HLEGs ethics guidelines.