



Project: 101094364 — ITHACA — HORIZON-CL2-2022-DEMOCRACY-01
 EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)
 REA.C – Future Society
 C.1 – Inclusive Society



ITHACA
 AI To Enhance Civic Participation

ITHACA
artificial Intelligence To enHance Civic pArticipation

D5.3 – ITHACA Conformity assessment mechanisms_v2 - report

Work Package: WP5 – Conformity assessments tools, policy recommendations and guidelines

Authors:	UPAT (Loi, Zachos, Moustakas), CERTH (Spiliotis, Chandrinos, Panou), SnP (Charikleia Eleni Nikolaou), UniGraz (Zangl, Bedek, Nussbaumer, Weichselgartner, Albert), SIMAVI (Dragota)
Status:	Final
Due Date:	31/12/2024
Version:	1.5
Submission Date:	20/12/2024
Dissemination Level:	PU - Public

Disclaimer:

This document is issued within the frame and for the purpose of the ITHACA project. This project has received funding from the European Union’s Horizon Europe Framework Programme under Grant Agreement No. 101094364. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

This document and its content are the property of the ITHACA Consortium. All rights relevant to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the ITHACA Consortium or the Partners detriment and are not to be disclosed externally without prior written consent from the ITHACA Partners. Each ITHACA Partner may use this document in conformity with the ITHACA Consortium Grant Agreement provisions.

(*) Dissemination level. - Public — fully open (automatically posted online)

Sensitive — limited under the conditions of the Grant Agreement

ITHACA Project Profile

Grant Agreement No.: 101094364

Acronym:	ITHACA
Title:	artificial Intelligence To enHance Civic pArticipation
URL:	https://www.ithaca-project.eu/
Start Date:	01/01/2023
Duration:	36 months

Partners

Short Name	Legal Name	Country
KT	KONNEKT ABLE TECHNOLOGIES LIMITED	IE
CERTH	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	EL
UPAT	PANEPISTIMIO PATRON	EL
RtF	RAISING THE FLOOR	BE
SnP	STAMADIANOS KAI SYNETAIROI DIKIGORIKI ETAIREIA	EL
UniGraz	UNIVERSITAET GRAZ	AT
MNLT	MNLT INNOVATIONS IKE	EL
SIMAVI	SOFTWARE IMAGINATION & VISION SRL	RO
PEDAL	PEDAL CONSULTING SRO	SK
BMA	AGENTIA METROPOLITANA PENTRU DEZVOLTARE DURABILA BRAȘOV ASOCIATIA	RO
MARTIN	MESTO MARTIN	SK

Document History

Version	Date	Author (Partner)	Remarks/Changes
0.1	14/10/2024	UPAT	Update of Structure
0.2	20/10/2024	UPAT, UniGraz, SnP	First Update of Chapters 1, 2 & 3
0.3	1/11/2024	UniGraz	Update of Chapter 1
0.4	5/11/2024	SnP	Update of Chapter 2 (Section 2.2)
0.5	7/11/2024	CERTH	Update of Chapter 3 (Section 3.3)
0.6	19/11/2024	UPAT	Update of Chapter 3 (Section 3.1.2)
0.7	6/12/2024	UPAT	Update of Chapter 3 (Section 3.4.1)
0.8	9/12/2024	UPAT	Update of Chapter 3 (Section 3.4.2)
0.9	10/12/2024	UPAT	Update of Chapter 3 (Section 3.4.3)
1.0	12/12/2024	UniGraz	Update of Chapter 1
1.1	15/12/2024	UPAT	Update of Chapter 4
1.2	17/12/2024	RtF	Reviewing and Commenting
1.3	18/12/2024	UniGraz	Reviewing and Commenting
1.4	19/12/2024	UPAT	Addressing Reviewer's Comments
1.5	20/12/2024	UPAT	Editing, Formatting and Submission to KT

Executive Summary

The ITHACA project develops a platform that aims to enhance civic participation, by also incorporating Artificial Intelligence (AI) methods and components. This platform should be as inclusive as possible, and the involved partners are striving to minimise the barriers and obstacles for citizens, especially for marginalised and vulnerable communities and individuals.

Regardless of the fact that there are a number of legal frameworks relating to the processing of personal data in general and the use of AI in particular, ethical handling of the data of all users, and especially of vulnerable and marginalised citizens, is considered essential. Ethics in AI encompasses a number of concepts - including fairness, security, privacy, transparency and explainability - which in most cases are closely interlinked. Thus, the ITHACA platform has to incorporate tools and components that ‘materialise’ these ethical concepts and principles into methods, metrics, algorithms and software for fair, secure and private AI, as well as to ensure that potential violations of these principles can be detected and remedied as quickly as possible.

This report extends its previous version (Deliverable 5.1) and focuses on important conformity assessment mechanisms in line with the principles of Fairness, Security, and Privacy. It defines qualitative and quantitative criteria for threats and risk evaluation stemming from the requirements for the ITHACA system and from the implementation of the AI systems for civic participation applications (see **Chapter 2**). The qualitative criteria are based on existing European legal regulations, in particular the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AI Act). The quantitative criteria are the result of a ‘translation’ of these qualitative criteria into computable metrics. **Chapter 3** deals with a set of three tools and one visual components: (i) *an AI fairness tool for fairness conformity assessment*, (ii) *a Privacy-Preserving Machine Learning (PPML) tool for privacy conformity in AI-based systems and big data for civic participation applications*, (iii) *an AI cybersecurity tool for security risk and threat detection*, as well as (iv) *a visual component for different stakeholders of the ITHACA platform (in particular a moderator) to efficiently control the data and to allow for a visual inspection and conflictive event detection*. Compared to the previous version (Deliverable 5.1), the (i) *AI fairness tool* has been extended by an *explainability mechanism* to render its decision-making process transparent to the user, for the (ii) *AI cybersecurity tool* an API has been implemented to facilitate the integration of an unsafe code detection tool (ModelScan) into deep learning models, such as the toxicity detection tool, and with regards to (iii) the visual component, different interfaces have been developed for two different stakeholders, users and moderators. Finally, **Chapter 4** summarizes the main updates and further developments since the submission of Deliverable 5.1. But first, **Chapter 1** introduces with the purpose and scope of this report in more detail, outlines the relations to other ITHACA deliverables, work-packages and tasks, and provides an overview on the structure and contents of the remaining document.

Table of Content

Executive Summary.....	5
Table of Content.....	6
List of Figures.....	8
List of Tables.....	9
List of Abbreviations.....	10
1 Introduction	11
1.1 Purpose and Scope.....	11
1.2 Relations to other Deliverables, WPs and Tasks.....	12
1.3 Structure of the Document.....	13
2 Qualitative and Quantitative Criteria	15
2.1 Risks and Threats.....	15
2.2 Qualitative Criteria.....	16
2.2.1 AI Fairness Tool.....	17
2.2.2 PPML Tool.....	22
2.2.3 AI Cybersecurity Tool.....	23
2.3 Quantitative Criteria.....	25
2.3.1 AI Fairness.....	26
2.3.2 Privacy Preserving Machine Learning.....	27
2.3.3 AI Cybersecurity.....	28
3 Evaluation Tools	30
3.1 AI Fairness Tool.....	30
3.1.1 Evaluation of an AI Toxicity Detection Model using the AI Fairness Tool	30
3.1.2 Explainability component	31
3.2 Privacy Preserving Machine Learning Tool.....	32
3.3 AI Cybersecurity Tool	33
3.4 Visual Component	36
3.4.1 Moderator Interface	36
3.4.2 User Interface	37
3.4.3 Accessibility Features	38
4 Conclusion	40
5 References	42

List of Figures

Figure 1: Visualization of the metrics for assessing the level of privacy of the PPML tool, namely, Area Under Curve, Attacker Advantage and Positive Predictive Value..... 33

Figure 2: An example of operation of ModelScan. 34

Figure 3: The steps of the scanning pipeline for machine learning models. 35

Figure 4: The prototype version of the Moderator’s Visual Component. 37

Figure 5: The prototype version of the User’s Visual Component. 38

List of Tables

Table 1: Definitions used throughout AI Fairness Sections

25

List of Abbreviations

Abbreviation /acronym	Description
AI	Artificial Intelligence
AUC	Area Under Curve
DP	Differential Privacy
DSA	Digital Services Act
EU	European Union
GDPR	General Data Protection Regulation
ML	Machine Learning
MLOps	Machine Learning Operations
NIS	Network and Information Systems
NLP	Natural Language Processing
PPML	Privacy Preserving Machine Learning
PPV	Positive Predictive Value
RCE	Remote Code Execution
TFEU	Treaty on the Functioning of the European Union
WP	Work Package
WCAG	Web Content Accessibility Guidelines
UI	User Interface

1 Introduction

This introductory Chapter briefly outlines the purpose and scope of this deliverable, and how it relates to other deliverables (D), work packages (WP) and tasks (T), in the sense that it receives inputs and delivers outputs. This document concludes with a summary on this deliverable's structure and its Chapters and Sections.

1.1 Purpose and Scope

This report is the third deliverable in WP5 (*'Conformity assessments tools, policy recommendations and guidelines'*) of the ITHACA project, and represents the updated version of D5.1 (*'ITHACA Conformity assessment mechanisms_v1 - report'*). It primarily focuses on conformity assessment mechanisms, particularly qualitative and quantitative criteria for evaluating threats and risks related to the ethical principles (in particular Fairness, Security and Privacy) underlying the platform. In consequence, these conformity assessment mechanisms build the foundation for the technical development of a set of tools that ensure and facilitate the above mentioned principles. The deliverable directly represents the results of the research and development activities carried out in the course of T5.1 (*'Fairness, Security and Privacy conformity assessment mechanisms'*). Its complementary 'technical counterpart', is represented by D5.4.

The main aim of Task 5.1 is to provide three tools that are applied alongside AI-based civic participation systems to evaluate the latter in terms of Fairness, Security and Privacy. More specifically, these tools include:

- An *AI Fairness tool* whose purpose is to conform with the fairness principle, to promote equality, inclusivity and oppose discrimination as well as render the evaluated AI system more trustworthy. In order to do so, the tool employs objective fairness metrics such as treatment equality and disparate impact, to evaluate whether an AI toxic censoring model infers a correlation between toxic speech usage and the group that the author of the post belongs to. In other words, it checks whether the toxicity detection model associates people from vulnerable groups with offensive language with a higher probability, thus, being biased and unfair. The fairness metrics were selected so as to comply with the respective legal requirements, while the toxic language detection functionality was deemed especially desirable in a civic engagement platform, according to citizen, civic workers and AI expert workers during WP2.
- A *Privacy Preserving Machine Learning tool* to employ data concealment techniques to safeguard user privacy. In particular, this first version of the PPML tool utilizes a Differential Privacy method to perturb sensitive user data by adding random noise, so that it would not be possible for a potentially malicious actor to infer any personal, potentially identifying information in case of a potential attack via specifically engineered inputs to the model. Such attacks attempt to infer whether a specific data sample belongs in the training dataset of an AI model. Attacks such as those, pose a threat to the privacy of data used for training a model and let alone in the case of this data stemming from a civic participation platform where some personal information is shared among the users.
- An *AI Cybersecurity tool* to detect possible security breaches and threats within the AI models. An open-source tool named ModelScan (Protect AI, 2023) that scans for malicious code that may impose security vulnerabilities in AI systems is proposed as the AI Cybersecurity tool. ModelScan focuses on detecting malicious injection during serialization (save) and deserialization (load) processes of the model. These stages can be exploited via Remote Code Execution (RCE) attacks. RCE allows adversaries to run external code into the platform's backend. This makes RCE threats extremely severe because it may

allow complete control of the platform. ModelScan, aims to mitigate this threat and to further protect the system from adversaries primarily seeking to collect sensitive data, incorporating an extra defense mechanism against cyber threats.

- The *Visual Component* is intended for a responsible party (i.e. human moderator) to access and monitor the three above-mentioned tools and, thus, have control and insight over the available data (which could be considered ‘big data’ after the platform is operational for a longer period of time) and events arising from the function of the evaluation tools mentioned above. There are two interfaces for different user-roles: The ‘moderator interface’ visualizes the results from the AI Fairness explainability mechanism and integrates results from both PPML and the AI Cybersecurity Tool, and the ‘end-user interface’ that provides users with weekly/monthly/lifetime statistics (e.g., regarding fairness, privacy etc.) and platform statistics, and visualizes the status of the three tools mentioned above (green, orange, red indications of the operational status of each tool).

The Visual Components (i.e. the one addressing the moderator as well as the one for end-user) will, once integrated into the platform inherit all its accessibility properties (e.g. color contrast option, voice navigation, image-to-text options, etc.) and will be easily accessible via a clearly visible button on the bottom right of the screen as is commonly the case with “help” elements in different tools.

1.2 Relations to other Deliverables, WPs and Tasks

As briefly outlined above, the direct input that resulted in this D5.3 is Task 5.1. However, Task 5.1 itself received *inputs* from several other tasks and deliverables of the ITHACA project, in particular:

- D1.1 (*‘Study on good practices of citizen engagement and democracy in AI applications’*)
This report collected, analysed and evaluated existing platforms that aim to facilitate and / or enable participatory democratic principles and collaborative decision making. The evaluation criteria of Fairness, Accountability, Privacy and Security. Based on the evaluation of all criteria, best practices have been identified. The outcome of this State-of-the-Art (SOTA) analysis of existing platforms built the foundation for the conceptualization as described in the final Chapter of D1.3 (Chapter 6).
- D1.3 (*‘Trustworthy AI compliance practices, assessment and conceptualization’*)
As mentioned above, on the one hand, D1.3 received input from D1.1 on existing platforms as well as their strengths and weaknesses, which provided the baseline for the ITHACA project to go beyond the SOTA w.r.t. a set of criteria and mechanisms. On the other hand, D1.3 (Chapter 4) analyzed existing compliance- and trustworthy AI systems (Section 4.2) as well as Privacy-Preserving Machine Learning Tools (Section 4.3). The analysis of these existing approaches and systems built the theoretical basis for the ITHACA components outlined in the Sections 3.1 to 3.3 in this D5.3. Finally, D1.3 also described and analyzed the SOTA of existing AI metrics on trustworthiness and ethical concepts, including metrics on Fairness, Privacy, Transparency and Explainability in Section 5 of D1.3. This encompassing analysis and comparison between quantifiable metrics provided an excellent starting point for the quantitative criteria to be incorporated as described in Section 2.3 of this D5.3.
- D2.1 (*‘Report on Citizen Jury Process and User requirements’*)
In the course of WP2, in particular in T2.4, an encompassing set of user requirements from a range of stakeholders have been gathered, analysed and condensed into use cases. The stakeholders were (i) end users from the AI Citizens’ Juries, (ii) municipality and city managers and administrators from both

ITHACA pilot cities (the Braşov metropolitan area in România and the City of Martin in Slovakia), as well as (iii) AI experts external to the ITHACA projects with diverse professional backgrounds (ethics, sociology, psychology, computer science, data science, etc.). These requirements and the associated use cases impacted the conceptualization, design and implementation of the conformity assessment mechanisms described in this D5.3.

- D5.1 (*'ITHACA Conformity assessment mechanisms_v1 - report'*) and D5.2 (*'ITHACA Conformity assessment mechanisms_v1 - prototype'*)

The outcomes of Task 5.1 are related to 4 Deliverables (D5.1 to D5.4), whereas this set can be seen as two pairs of complementary counterparts: D5.1 and D5.3 describe the *conformity assessments mechanisms and tools* from a conceptual point of view (incl. their purpose, legal context, technical design, and application for the ITHACA platform), whereas D5.2 and D5.4 focus on the actual technical implementation as prototypes. D5.1 and D5.2 cover the 1st version, whereas D5.3 and D5.4 will cover the final, updated version of the *conformity assessments mechanisms and tools*.

- D8.2 (*'Personal Data Protection Handbook'*)

This report outlines the main obligations and legislative rules regarding data protection at the EU level, with a main focus on the GDPR. This analysis is related to the qualitative criteria and legal requirements as described in Section 2.2 in this D5.3.

- D8.3 (*'Artificial Intelligence and Ethical Issues'*)

This report outlines the main obligations and legislative rules related to AI, such as the AI Act. As for D8.2, the according outcomes are related to the qualitative criteria and legal requirements as described in Section 2.2 in this D5.3.

The results of this D5.3 constitute an *output for* the ITHACA platform as such, as well as WP4 (*'Pilots' implementation & evaluation'*) in which the components will be tested and evaluated with end-users from the two ITHACA pilot cities. Finally, as the ITHACA consortium aims to disseminate scientific publications, it offers a range of opportunities for dissemination activities and cooperation with other similar research projects and initiatives.

1.3 Structure of the Document

From a bird's-eye view, this Deliverable is structured as follows (direct quotes from the Task 5.1 description as in the Grant Agreement are in *italic*):

- **Chapter 1** outlines the purpose and scope of deliverable D5.3, and how it relates to other deliverables (D), work packages (WP) and tasks (T).
- **Chapter 2** describes the *'risks and threats arising from the development phase of the AI systems for civic engagement applications'* (Section 2.1) and defines *qualitative* (Section 2.2) and *quantitative criteria* (Section 2.3).
- **Chapter 3** specifies *'three scalable tools based on [...] visual analytic technologies, for the identification of security risks and threats (AI cybersecurity tool [see Section 3.3]) for conformity with the Fairness principle (AI fairness tool [see section 3.1]) and for conformity with the Privacy principle (PPML tool [see Section 3.2]) in AI-based systems and big data for civic engagement applications'*. In addition, a visual component for efficient i) control of big data and ii) visual inspection and event detection is specified (see Section 3.4).

- **Chapter 4** summarizes the main updates and further developments since the submission of Deliverable 5.1.
- Finally, **Chapter 5** lists the references.

2 Qualitative and Quantitative Criteria

2.1 Risks and Threats

In the development of civic engagement platforms like ITHACA, addressing legal considerations related to risks and threats is crucial to ensuring that the platforms operate within a safe and legally compliant framework. These considerations primarily revolve around data protection, content moderation, non-discrimination, and cybersecurity. Each of these areas presents unique challenges that require meticulous attention to both legal mandates and ethical standards.

Civic participation platforms are fundamentally designed to enhance democratic engagement by providing, among others, a digital space where a wide range of citizens (including people from vulnerable groups) can access, discuss, deliberate, and engage with public issues. These platforms leverage technology to facilitate a broader range of input on public policies and community decisions, effectively lowering barriers to participation that might exist in more traditional settings of civic engagement. This democratisation of participation relies heavily on the ability of the platform to maintain an environment where users feel safe and valued, free from harassment or discrimination. The nature of these platforms is such that they directly influence the quality of democratic discourse by shaping how information is shared and discussed among the citizenry (Sunstein, 2001). The importance of maintaining a robust platform that supports these democratic ideals cannot be overstated, as it directly impacts the efficacy of civic engagement and the overall health of democratic processes.

The relationship between AI technologies and democracy within these platforms presents both opportunities and challenges. On one hand, AI can greatly enhance the scalability of civic engagement platforms, allowing them to process large volumes of data and interactions more efficiently, thereby supporting more nuanced and widespread participation. On the other hand, if not carefully managed, AI can introduce biases or manipulate discussions, potentially skewing public opinion or marginalising certain groups (Helbing et al., 2019; Woolley and Howard, 2016). These risks and potential harms highlight the necessity for transparency in AI algorithms and the rigorous testing of AI systems to ensure they uphold the principles of fairness and equality fundamental to democratic participation. Moreover, the use of AI in these contexts must be continuously monitored and adjusted based on feedback from users and changes in societal norms to ensure that the technology serves to support, rather than undermine, democratic values. In the context of Task 5.1, three scalable tools will be developed to address critical considerations related to the development and use of AI platforms in the field of civic participation.

Firstly, data protection and privacy are paramount under the General Data Protection Regulation, which mandates that personal data must be processed lawfully, fairly, and transparently (European Parliament and Council of the European Union, 2016). Civic platforms must implement systems that protect user data from unauthorized access, data breaches and unlawful process of data while ensuring that the data collection and processing activities are clear to users (transparency principle). The Privacy Preserving Machine Learning technologies, by introducing techniques such as noise addition to pseudonymize data, directly addresses these privacy concerns and aligns with Article 32 of the GDPR, which requires the implementation of appropriate technical and organizational measures to ensure a level of security appropriate to the risk (Regulation (EU) 2016/679, Article 32; Mayer-Schönberger and Cukier, 2013).

Content moderation is another critical area, as platforms must balance the need to curb harmful speech without infringing on free expression, taking into account the provisions of the applicable EU legislation, such as the

Digital Services Act (hereinafter: “DSA”). The detection of toxic speech is a response to this requirement, aiming to ensure that content moderation algorithms do not discriminate against certain groups or suppress legitimate discourse (Gillespie, 2020). Non-discrimination is a fundamental principle under the Charter of Fundamental Rights of the European Union, which prohibits any discrimination based on grounds such as race, gender, or religious belief (Charter of Fundamental Rights of the European Union, 2000/C 364/01, Article 21). The necessity of an AI Fairness Tool is critical, as it plays a key role in assessing whether algorithms discriminate against specific user groups, thereby helping platforms comply with non-discrimination laws (Barocas and Selbst, 2016).

Finally, cybersecurity is a significant concern given the vulnerabilities that can be exploited in digital platforms. The development of the cybersecurity tool, aligns with the Directive on Security of Network and Information Systems (hereinafter: “NIS 2 Directive”), which mandates operators to implement appropriate security measures to manage risks posed by network and information systems (Schneier, 2018).

In the context of ITHACA platform Remote Code Execution was found to be a severe threat to AI models. RCE is a common type of cyber attack that allows an attacker to run arbitrary code on a target machine or server from a remote location. This code can perform a wide range of actions, from stealing data to creating new accounts with full user rights, depending on what the attacker embeds within the code and the permissions available on the target system (Yao, Y. et al., 2024). RCE vulnerabilities are particularly severe because they can be exploited to take complete control of the affected system without needing physical access (NIST AI-100-2 E2023).

This threat comes from a weakness in the loading and saving process of AI models, known as deserialization vulnerability. When a serialized Machine Learning (ML) model is loaded into an application, any malicious code embedded within the model during the serialization process can be executed automatically (Liu, Y. et al., 2019). To safeguard AI models throughout their use, stricter security measures are crucial to address this deserialization vulnerability.

While AI fairness and PPML are important aspects of AI development, it is crucial to emphasize that security breaches can significantly compromise these attributes. When a model is attacked, it may not only lose its fairness, leading to biased outcomes, but also its privacy-preserving capabilities, potentially resulting in the exposure of sensitive information (Chan, L. et al., 2019).

In conclusion, based on the abovementioned legal considerations and challenges, qualitative and quantitative criteria aimed at proposing innovative methods for ensuring compliance and fostering trustworthy development of AI civic engagement platforms at the EU level, are being developed. By aligning with legal standards and ethical considerations, these tools not only mitigate legal risks and threats associated with running civic platforms but also lay the groundwork for a comprehensive evaluation of how these criteria are implemented and assessed in real-world settings.

2.2 Qualitative Criteria

The development of qualitative criteria is essential for evaluating the potential threats posed by the deployment of AI tools in civic participation platforms. These criteria are vital to ensure the tools are effectively integrated, operate ethically, and comply with legal standards, thereby maintaining the integrity and functionality of the ITHACA platform. The effectiveness of AI tools hinges on their ability to deliver accurate and reliable outputs.

Additionally, compliance with the legal framework is another critical aspect. AI tools must ensure that ITHACA platform adheres to stringent legal requirements concerning data protection, privacy, and the prevention of discrimination. This includes aligning operations with comprehensive regulations such as GDPR, AI ACT, DSA, NIS2 Directive, and preparing for adaptability to accommodate future legal changes.

By laying out these qualitative criteria, the groundwork is set for identifying quantitative criteria that can offer more precise measures of the AI tools' performance and alignment with regulatory standards.

In this updated version of deliverable D5.1, namely D5.3, qualitative criteria remain unchanged since no new legislations concerning the compliance of AI systems in terms of fairness, transparency, privacy, security, data protection and data governance, and equity and inclusiveness, have been put into effect in the intervening time between these two documents. The latter also affects Section 2.3, where quantitative criteria are consistently upheld and adhered to throughout the development of the final versions of the evaluation tools.

2.2.1 AI Fairness Tool

In the dynamic interface between Artificial Intelligence and civic engagement, the development of tools like the fairness AI tool, which is instrumental in detecting toxic speech on platforms such as ITHACA, embodies a critical juncture of technology and social responsibility. The ITHACA Platform, aimed at enhancing civic participation, integrates this AI tool to foster a respectful and inclusive online environment. To navigate the complexities of such an integration effectively, it's imperative to ground the AI tool's development and deployment in a robust legal and ethical framework. The initial segment of our exploration delineates the legal requirements stemming from two pivotal regulatory pillars: the General Data Protection Regulation and the Artificial Intelligence Act (AI Act). These legal standards lay the groundwork for ensuring that AI technologies not only adhere to the highest levels of data protection and ethical conduct but also align with societal values and rights.

Moving beyond the legal imperatives, the subsequent part of our discourse shifts focus towards the qualitative criteria that define the operational essence of the fairness AI tool within the ITHACA Platform. This segment aims to translate specific legal mandates into practical, actionable guidelines, ensuring that the tool effectively identifies toxic speech while upholding the principles of fairness, transparency, and accountability intrinsic to civic participation. Through this bifocal approach—merging legal requirements with qualitative criteria—we endeavor to outline a comprehensive blueprint for the AI tool's contribution to the ITHACA Platform, underpinning its role not just as a technological solution but as a facilitator of healthy civic dialogue and engagement.

2.2.1.1 Legal Requirements

2.2.1.1.1 GDPR

Fairness Principle [Art. 5 par. 1 (a)]

Among GDPR various principles, fairness stands out as critical, especially when applied to the development and deployment of Artificial Intelligence systems, such as those designed to mitigate toxic speech. This qualitative criterion focuses on ensuring that AI tools not only comply with GDPR's fairness principle but also embody it in their operational essence, particularly in processing personal data in a way that is fair to the individuals concerned. The criterion encompasses the need for transparency, accountability, and equity in AI processes, directly impacting the design, implementation, and evaluation phases of AI tools.

Firstly, to meet the fairness principle under GDPR, AI systems, particularly those dealing with sensitive issues like toxic speech, must ensure transparent data processing activities. Transparency in this context means that individuals are fully informed about how their data is being used, the purpose of its use, and the implications of this usage. This is especially pertinent in AI systems where data processing can often be opaque or too complex

for the average user to understand. Fairness, therefore, requires that efforts be made to demystify AI operations, providing accessible explanations for non-technical users, thus enabling informed consent and participation in digital ecosystems (Information Commissioner's Office, 2020).

Furthermore, the fairness principle demands that AI systems are designed and operated in a manner that prevents discriminatory outcomes. This involves scrutinizing training data to ensure it is representative and free from biases that could harm vulnerable groups. The GDPR's emphasis on fairness necessitates ongoing vigilance to identify and rectify biases, a process that should be embedded within the AI development lifecycle (European Union, 2016).

The principle of Non-discrimination in EU Law

The aim of non-discrimination law is to allow all individuals an equal and fair chance to access opportunities available in a society. This means that individuals or groups of individuals which are in comparable situations should not be treated less favourably simply because of a particular characteristic such as their sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation.

The Treaty on the Functioning of the European Union (TFEU) prohibits discrimination on grounds of nationality. It also enables the Council of the European Union to take appropriate action to combat discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation. In this matter, the Council must act unanimously and after obtaining the European Parliament's consent. However, in the specific area of equal treatment and equal opportunities for men and women, the ordinary legislative procedure applies, which does not require unanimity but only qualified majority (Article 157 TFEU).

Discrimination on the grounds of nationality has always been forbidden by the European Union (EU) treaties, as has discrimination on the basis of sex in the context of employment. The other grounds of discrimination were mentioned for the first time in 1997, with the signature of the Treaty of Amsterdam.

In 2000, two directives were adopted:

- the employment equality directive (Directive 2000/78/EC), which prohibits discrimination on the basis of sexual orientation, religious belief, age and disability in the area of employment;
- the racial equality directive (Directive 2000/43/EC), which prohibits discrimination on the basis of race or ethnicity, again in the area of employment but also in areas such as education, social protection including social security and healthcare, social advantages and access to and supply of goods and services, including housing.

EU legislation also protects people against discrimination based on their sex in the above areas, except for education.

In 2009, the Treaty of Lisbon introduced a horizontal clause with a view to integrating the fight against discrimination into all EU policies and measures (Article 10 TFEU).

In practical terms, implementing the fairness principle within AI systems involves developing comprehensive impact assessments that consider the potential effects of AI decisions on individuals and groups, especially those that are vulnerable or marginalised. Regular reviews and updates to AI systems, in response to feedback and changing societal norms, are also essential to maintain fairness over time (European Commission, 2020).

Transparency Principle [Art. 5 par. 1 (a)]

Transparency in AI requires detailed disclosure regarding the AI system's design and decision-making processes. Users should be provided with comprehensible information about how the system identifies toxic speech, including the criteria it uses and the types of data it analyses. This goes beyond technical explanations to include accessible descriptions of the methodologies and logic applied, ensuring that even those without a background in data science can grasp how the system works. Such openness is essential not only for fostering user trust but also for complying with GDPR mandates that call for clear communication about the processing of personal data (European Parliament and Council of the European Union, 2016).

The implementation of transparency also entails providing users with avenues for recourse and feedback. A transparent AI system must include mechanisms that allow users to question decisions made by the AI, particularly those related to the identification and moderation of toxic speech. This not only adheres to GDPR's stipulations on the right to explanation but also enhances the system's accountability by making its operations subject to scrutiny and improvement based on user input (European Data Protection Board, 2020).

Meaningful information of data subjects [Art.14.2(g)]

Meaningful information implies that data subjects are fully informed about the processing of their data. This goes beyond mere notification of data collection; it requires a detailed explanation of the purposes of data processing, the types of data being processed, and the entities involved in this processing. For AI systems moderating toxic speech, this means clearly articulating how individuals' data—whether textual content, metadata, or user interactions—is used to train algorithms or make content moderation decisions. Such transparency ensures that data subjects understand the scope and implications of their data's use, aligning with GDPR's mandates for clear and precise information provision (Article 13, GDPR).

Moreover, the requirement of meaningful information encompasses the necessity for data subjects to understand the rationale behind AI-driven decisions. This is particularly pertinent in scenarios where AI systems determine what constitutes toxic speech and decide on the appropriate actions to take, such as content removal or user bans. The ITHACA +platform must strive to demystify the AI's decision-making processes, offering insights into how algorithms interpret data and arrive at conclusions.

2.2.1.1.2 AI ACT

Data and Data Governance [Article 10 par. 2 (f) and (fa)]

Proactive Bias Management in High-Risk AI Systems

- **Systematic Bias Identification**

This legal requirement demands a proactive and thorough approach to identifying potential biases in training, validation, and testing datasets, as well as within the algorithms that drive decision-making processes in AI systems. This step is pivotal, recognizing that biases can manifest in multifaceted ways, influencing the system's behavior and decisions, potentially leading to adverse outcomes for certain individuals or groups. The identification process must be comprehensive, involving statistical analyses to uncover hidden biases and qualitative assessments to understand their context and implications. Such diligence ensures that biases are not merely detected but are understood in terms of their potential impact on the health, safety, and rights of individuals.

- **Development and Application of Mitigation Techniques**

Following the identification of biases, the next essential step within this requirement is the development and application of robust mitigation techniques. This involves revising data collection and preparation

methodologies, enhancing algorithmic transparency, and incorporating fairness-oriented machine learning approaches.

Transparency and provision of information to deployers [Art. 13 par. 3 (iv)]

Ensuring transparency and understanding for individuals

It's crucial to adhere to the principles outlined in Article 13, paragraph 3, subsection iv of the AI Act. This subsection mandates clear communication about the logic of AI decisions, especially those that have significant consequences for individuals. For vulnerable populations, understanding how AI moderates toxic speech and protects their rights is paramount. This criterion, therefore, focuses on ensuring that the AI's decision-making processes are transparent, understandable, and accessible to these groups.

Accessibility and clarity of AI moderation Logic

This requirement underlines that the AI system's moderation logic, particularly in identifying and responding to toxic speech, be communicated in a manner that is both accessible and comprehensible to all users, including vulnerable populations. Information should be provided in multiple formats to accommodate different accessibility needs. This includes clear explanations of how the AI identifies toxic speech, the standards it uses, and the actions it may take in response.

Human Oversight (Article 14)

In the realm of civic participation platforms, the imperative for AI fairness is underscored by the emerging legislative frameworks, particularly through the lens of the Human Oversight principle encapsulated within the AI Act. This principle mandates the incorporation of human judgement in the AI decision-making process, ensuring that automated systems do not solely dictate outcomes that impact civic engagement and participation. By integrating human oversight, AI platforms can effectively mitigate biases inherent in AI algorithms, promoting a more representative and fair digital civic space (Kullmann and Cefaliello, 2021).

The implementation of AI fairness tools in civic participation platforms necessitates a framework that aligns with the Human Oversight principle, advocating for a balanced interplay between automated processes and human discretion. Such a framework should prioritize transparency, accountability, and the ability to challenge AI-driven decisions. In practice, this involves the establishment of oversight mechanisms wherein human moderators or oversight bodies review and potentially override AI decisions that could influence civic participation. These measures ensure that AI systems act in the public interest, respecting the diversity of public opinion and protecting against the amplification of biases (Pinto, 2021).

The transition from the foundational legal requirements outlined in the GDPR and the AI Act to the establishment of qualitative criteria for AI fairness is a critical juncture in the discourse on ethical AI. This passage involves translating the abstract principles of law into actionable guidelines that can be directly applied in the development and operation of the AI Fairness tool.

2.2.1.2 Identification of Qualitative Criteria

A. Equity and Inclusiveness

The qualitative criterion of equity and inclusiveness is central to addressing and mitigating disparate impacts and minimising statistical parity differences. This aligns with both the Fairness Principle of the GDPR [Art. 5 par. 1 (a)] the Non-discrimination principle of EU Law as well as the Bias identification and mitigation requirement under the proposed AI Act. Quantitatively, reducing disparate impact ratios and achieving closer statistical parity differences would indicate compliance with these legal mandates, demonstrating the AI system's commitment to equitable treatment across all user demographics.

B. Transparency and Explainability

Ensuring transparency and explainability in AI systems is fundamental to aligning with the Transparency Principle of the GDPR [Art. 5 par. 1 (a)] and the Transparency and provision of information to deployers outlined in the AI Act [Art. 13 par. 3 (iv)]. An increase in the accuracy of correct positive and negative predictions serves as a quantitative measure of the system's transparency, highlighting the AI's capacity to deliver clear and understandable decisions. This improvement not only showcases the AI system's compliance with legal mandates for transparent communication about its data processing and decision-making process but also builds trust among users and stakeholders by demonstrating the system's reliability and accountability.

Furthermore, the AI tool should offer straightforward and comprehensible information about the criteria and logic used to determine that a post contains toxic speech. While it might not be feasible to detail the intricate workings of a complex algorithm fully, providing a simplified example that clearly illustrates the particular segment of speech that was flagged as toxic can significantly aid user understanding. This approach empowers users by making it easier for them to challenge or question the AI's decisions, in line with their rights under the GDPR and the provisions of Article 14 of the AI Act. This level of transparency is crucial for enabling users to lodge objections or seek clarification from human overseers, thereby reinforcing the principles of fairness and accountability in AI applications.

C. Upholding Human Oversight and the Right Against Solely Automated Decisions in AI Systems

The incorporation of a clear explicative mechanism within AI systems, elucidating the process behind the identification of toxic text, aligns precisely with the principles of Human Oversight and individuals' rights to contest solely automated decisions, as enshrined in the General Data Protection Regulation [Art. 22] and echoed in the Artificial Intelligence Act (AI Act). By underlying the part of the text detected as toxic, the system not only complies with the GDPR's requirements for transparency and the provision for human intervention [GDPR Art. 5(1)(a), Art. 22] but also upholds the mandates of the AI Act, specifically regarding transparency and human oversight [AI Act Art. 14, Art. 13 par. 3(iv)]. This methodology not only fortifies trust and accountability but ensures that AI-mediated decisions, especially those impacting individuals' freedom of expression and rights, maintain transparency and are open to reassessment and amendment by human moderators.

By embedding these qualitative criteria within the AI development framework, and linking them to specific quantitative metrics and legal provisions of the GDPR and the proposed AI Act, AI systems are better positioned to meet current and future regulatory requirements. This holistic approach fosters the development of AI tools that are not only legally compliant but also ethically aligned and socially responsible, enhancing civic participation and ensuring a fair and inclusive digital society.

2.2.2 PPML Tool

2.2.2.1 *Legal Requirements*

2.2.2.1.1 GDPR

A. Data Minimization and Purpose Limitation [Art. 5 par. 1 (b) & (c) GDPR]

Under the GDPR, PPML tools on the ITHACA platform must strictly adhere to the principles of data minimization and purpose limitation. Data minimization ensures that the collection and processing of personal data are limited to what is absolutely necessary for specified purposes. In developing a PPML tool, this means avoiding the collection and processing of unnecessary data that could make individuals identifiable or easily identifiable, thereby reducing the risk of data breaches and enhancing privacy protection (GDPR, Art. 5(1)(c); Alaya et al., 2020).

Purpose limitation mandates that personal data must be collected for specified, explicit, and legitimate purposes and not further processed in a manner incompatible with those purposes. PPML tools should ensure that data is only used for these predefined purposes, and any further processing should occur in a way that does not make individuals identifiable or easily identifiable, thereby upholding the purpose limitation principle (GDPR, Art. 5(1)(b); Xu et al., 2021).

Together, these principles play a critical role in safeguarding user privacy by preventing the misuse of personal data. They ensure that PPML tools on the ITHACA platform operate within the bounds of legal compliance, maintaining user trust and data integrity (Schneier, 2018).

B. Integrity and Confidentiality [Art. 5 par. 1 (f) GDPR] and Technical and Organizational Measures [Art. 32 GDPR]

The GDPR emphasizes the necessity of maintaining the integrity and confidentiality of personal data through robust technical and organizational measures. PPML tools shall implement advanced data protection methods such as pseudonymization, anonymization, and secure processing to comply with these requirements. These methods ensure that personal data is shielded from unauthorized access and manipulation, significantly reducing the risk of data breaches and enhancing privacy protection (GDPR, Art. 5(1)(f)).

In addition to these measures, Article 32 of the GDPR mandates that appropriate technical and organizational measures proportional to risk levels be utilized, incorporating state-of-the-art technology for data protection. PPML tools are essential in fulfilling these requirements by enhancing the security of the platform.

2.2.2.1.2 AI ACT

C. Data Governance [Art. 10 AI Act]

The AI Act specifies stringent requirements for data governance, particularly for high-risk AI systems that involve the processing of special categories of personal data. Article 10 emphasizes that training, validation, and testing data sets must adhere to high standards of data governance and management practices appropriate for the intended purpose.

To ensure the integrity and confidentiality of personal data, the AI Act mandates that these data sets are handled with **state-of-the-art security and privacy-preserving measures, including pseudonymization**. This ensures

that personal data are protected against unauthorized access and misuse, maintaining high standards of data protection throughout the AI system's life cycle.

2.2.2.2 *Identification of Qualitative Criteria*

A. Compliance with Advanced Data Protection Measures [AI Act Art. 10]

The AI Act mandates advanced data protection measures for high-risk AI systems, particularly those processing special categories of personal data. PPML tools must incorporate technical restrictions on data reuse **and apply privacy-enhancing technologies** like pseudonymization. These measures ensure the confidentiality and integrity of sensitive data, protecting it from unauthorized access and breaches. Implementing such robust data governance practices is essential for compliance and safeguarding user privacy throughout the data lifecycle.

Differential Privacy (DP) is an advanced technique that could be incorporated in the PPML tool to meet these stringent privacy requirements. **By adding random noise to data, DP techniques like gradient clipping and noise injection help protect individual privacy during machine learning processes.** This should ensure that the data cannot be easily traced back to specific individuals, significantly reducing the risk of re-identification.

Integrating Differential Privacy with advanced data protection measures allows PPML tools to balance data utility and privacy.

B. Balancing Data Utility with Privacy Preservation

Balancing data utility with privacy preservation is a critical criterion for the effective implementation of the PPML tool. Ensuring that data remains useful for machine learning while protecting individual privacy requires sophisticated techniques, such as the previously mentioned DP approach, which allows for accurate model training and analysis in a privacy friendly way. This balance ensures that the models are both effective and compliant with privacy regulations.

Maintaining high data utility means that the processed data retains its relevance and accuracy for the intended machine learning tasks, ensuring compliance with GDPR's requirement for data accuracy.

The abovementioned qualitative criteria ensure that the PPML tool on the ITHACA platform not only comply with the legal frameworks of the GDPR and the AI Act but also maintain high standards of data protection and privacy, thereby supporting the platform's objectives and enhancing user trust.

2.2.3 **AI Cybersecurity Tool**

2.2.3.1 *Legal Requirements*

2.2.3.1.1 GDPR

Technical and Organisational Measures (GDPR, Article 32)

Data controllers must implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk. These measures should be designed to protect personal data against unauthorized access, accidental loss, destruction, or damage, thereby ensuring confidentiality, integrity, and availability of processing

systems and services throughout their lifecycle. This involves the use of encryption, pseudonymization, and other advanced security technologies to safeguard personal data.

2.2.3.1.2 AI ACT

Accuracy, Robustness, and Cybersecurity Compliance (AI Act, Article 15)

High-risk AI systems must be designed and developed to achieve an appropriate level of accuracy, robustness, and cybersecurity throughout their lifecycle. High-risk AI systems must be resilient against attempts by unauthorized third parties to alter their use, outputs, or performance by exploiting system vulnerabilities. This includes implementing measures to prevent, detect, respond to, resolve, and control attacks such as data poisoning, model poisoning, adversarial examples, and confidentiality attacks. Systems that continue to learn post-deployment must be designed to minimize the risk of biased outputs influencing future operations, with appropriate feedback loop mitigation measures in place.

2.2.3.1.3 NIS 2 Directive

Cybersecurity Risk Management Measures and Reporting Obligations (NIS 2 Directive, Articles 20-25)

The NIS 2 Directive mandates that entities implement comprehensive cybersecurity risk management measures, which include protecting network and information systems from cybersecurity threats and incidents. This involves identifying and managing risks, ensuring continuity of services, and reporting significant incidents to relevant authorities. A cybersecurity tool could contribute to these measures by scanning machine learning models for vulnerabilities that could be exploited, thereby enhancing the overall cybersecurity posture of the ITHACA platform in compliance with the NIS 2 Directive's requirements (Directive (EU) 2022/2555, Articles 20-25).

2.2.3.2 *Identification of Qualitative Criteria*

The identification of qualitative criteria for cybersecurity tools is directly informed by the requirements described by GDPR, AI Act and NIS 2 Directive. (Sarker, I. H. et al., 2021).

A. Robust Mitigation against Remote Code Execution

AI models are particularly vulnerable to serialization vulnerabilities during the loading process. Serialization vulnerabilities can allow attackers to execute arbitrary code on a machine that processes serialized data, which poses a significant risk especially when models are loaded from untrusted sources or over a network. These stages are susceptible to infiltration by adversaries who can embed malicious code into the models. A dedicated cybersecurity tool is required to scan and monitor these stages effectively, ensuring that any such injections are identified and neutralized promptly. The proposed tool should include capabilities to thoroughly analyze the model and operations involved in the serialization and deserialization processes. By detecting anomalies or patterns indicative of RCE attempts, the tool should provide alerts and protective measures to prevent the execution of unauthorized code.

B. Scalability and Ease of Use for Integration

For integration into real-world machine learning workflows, a scanning tool must offer high-performance scanning that does not become a bottleneck. This includes quick scan times even as model sizes and complexities increase, ensuring that the tool can scale with the demands of modern AI development environments. Integrating security tools into existing workflows should be straightforward to encourage widespread adoption and continuous

use. This tool should feature easy setup processes, clear documentation and a user-friendly interface in order to accommodate both novice and experienced developers. A comprehensive scanning tool must support a wide range of model formats and be compatible with numerous machine and deep learning frameworks. An example of a suite of machine / deep learning frameworks could be Tensorflow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), Scikit-learn (Pedregosa et al., 2011), etc. Tensorflow is an open-source framework for building and training deep learning models. Similarly PyTorch is another open-source framework, favored for its dynamic computation and ease of use in research. Scikit-learn is a library in Python for machine learning that provides simple and efficient tools for data analysis and machine learning, with a focus on ease of use during integration.

2.3 Quantitative Criteria

Based on the previously described qualitative criteria as well as the review of AI evaluation approaches and metrics in D1.3, a conclusion was made regarding which methods concerning AI Fairness, PPML and AI Cybersecurity comply with the legal requirements as specified in Section 2.2.

The Fairness tool developed for the purposes of the ITHACA platform, was designed to be in accordance with the Qualitative requirements outlined in Section 2.2, with a strong focus on the practices suggested by EU laws such as the AI Act and GDPR as described in Section 2.2.1.2. Specifically, the metrics used to evaluate the developed AI models, which essentially provide a mathematical description of the Quantitative criteria with the goal of making the system in its entirety more Trustworthy, are matched with the Qualitative criteria.

Accordingly, for the PPML tool, the Differential Privacy approach will be used, which applies arbitrary modifications such as the addition of random noise to the data through differential procedures (Gaussian, Laplace, exponential, etc.) to a dataset such that if an individual has access to the dataset's entries, they will not be able to infer any personalised sensitive information from it (Xu, 2021, Zapechnikov, 2020). At this point, It is worth mentioning that until the time this deliverable is written, the partners are not aware whether collaborative ML learning methods may apply to the ITHACA platform, thus making multi-party computation and federating learning less suitable. Having stated that, the main reasons why DP was preferred for the development of the PPML tool, are analysed below:

- According to the analysis of PPML approaches in D1.3, Differential Privacy has many advantages compared to other common methods, since:
 - Anonymization, which involves the concealment or removal of sensitive features from a dataset, is vulnerable to de-anonymization attacks, while an attacker can also infer personal information from the rest of the dataset since no encryption/modification method is applied to keep the data private (Xu, 2021).
 - Multi-party computation and federated learning consist of collaborative ML methods, meaning that multiple contributors to an AI-based model can train it and produce inferences locally using data kept private on their systems. Furthermore, multi-party computation suffers from communication and computation overhead, while federated learning is vulnerable to privacy leakages (Xu et al., 2021).
 - Homomorphic encryption consists of a cryptographic approach that allows complex computations on encrypted data. However, this method has several limitations; limited

functionality, high computational complexity, slowness, and many more as discussed in (Alaya, 2020).

- Moreover, based on legal guidelines as described in Section 2.2.2:
 - DP is in line with the GDPR Art. 5 par.1 (f) “Integrity and Confidentiality” (qualitative criteria A), which states that PPML should perform robust data concealment methods to maintain the privacy of data.
 - DP complies with Art. 32 of GDPR and Art.10 of AI Act (qualitative criteria B), where apart from the appliance of sophisticated privacy-proof methods that DP entails, there are local differential privacy approaches, which allow data owners to maintain the confidentiality of personalised data locally before sharing them, (Arachchige et al., 2020). Therefore, local DP methods pose both technical restrictions on the reuse (i.e. for training/evaluating AI models) of special categories of data since they are kept private locally in the systems of the owners. Organising the data in such a way, leads to technical precautions against de-anonymization and other attacks.
 - As for qualitative criteria C, PPML methods should retain the privacy of data, while not hindering the performance of the model trained/validated/tested upon them. It is worth mentioning that DP techniques suffer from loss of model utility (e.g. reduced accuracy) according to (Xu, 2021). However, recent local DP approaches, such as *LATENT* (Arachchige et al., 2020), tackle this limitation as deduced through experiments on well-known image datasets like CIFAR-10 (Krizhevsky et al., 2009) and MNIST (LeCun and Cortes, 2010). Moreover, global DP approaches, where both the data and DP method that perturbs these data, reside in a server, also perform well with accuracy reaching ~93% on MNIST (Phan et al., 2017) as shown in D1.3 (T1.2).

In selecting a cybersecurity tool for the ITHACA platform, it was imperative to choose a solution that not only meets technical requirements but also adheres to stringent legal standards outlined in the GDPR, AI Act, and NIS 2 Directive. The selected AI cybersecurity tool offers comprehensive compatibility with various machine learning frameworks and supports multiple model formats, ensuring seamless integration into diverse technological environments. This flexibility is crucial for accommodating the wide-ranging needs of modern AI developments. By meeting these technical and legal criteria, the chosen tool not only enhances the security defences of AI systems on the platform but also ensures compliance with critical EU regulations.

2.3.1 AI Fairness

In the pursuit of responsible AI development, ensuring fairness in model outputs is paramount. This Section explores a range of quantitative AI Fairness metrics designed to detect and mitigate bias. We will dissect metrics such as Disparate Impact, which unveils discrepancies in model outcomes for different groups. Statistical Parity Difference will be scrutinised, revealing the extent to which a model deviates from equal selection rates. We will then delve into Equal Opportunity Difference, a metric that sheds light on disparities in receiving positive outcomes across various groups. Furthermore, Between Group Generalised Entropy Error will be investigated, a measure that quantifies the difference in model accuracy between groups. Finally, the concept of Treatment Equality will be introduced, ensuring that the model treats all individuals similarly, irrespective of their attributes. Table 1 presents the definitions of common terms used when describing the Fairness metrics employed in the following Sections.

Table 1: Definitions used throughout AI Fairness Sections

Term	Meaning
Positive Prediction	A post that is classified by the algorithm as not-toxic. Includes Correct Positive Predictions as well as False Positive Predictions.
Correct Positive Prediction	A post that is classified by the algorithm as non-toxic is actually <i>non-toxic</i>
False Positive Prediction	A post that is classified by the algorithm as non-toxic is actually <i>toxic</i>
Negative Prediction	A post that is classified by the algorithm as toxic. Includes Correct Negative Predictions as well as False Negative Predictions.
Correct Negative Prediction	A post that is classified by the algorithm as toxic is actually <i>toxic</i>
False Negative Prediction	A post that is classified by the algorithm as toxic is actually <i>non-toxic</i>

Following we present a short description of the metrics that were selected to be employed by the AI Fairness tool, always in accordance with the legal requirements described in Section 2.2.1.2. These metrics were analyzed and reviewed in deliverable D1.3.

Disparate Impact: The *ratio* of the probability of **positive predictions** for one group over the proportion of positive predictions for another. Generally, if the unprivileged group receives a positive outcome less than 80% of their proportion of the privileged group, this is a disparate impact violation.

Statistical Parity Difference: The *difference* between the probability of unprivileged group getting a **positive prediction** versus the probability of privileged group getting a **positive prediction**. Generally a difference more than 10% leads to a statistical parity difference violation.

Equal Opportunity Difference: A relaxed version of equality of opportunity. The difference in **correct positive** predictions between the unprivileged and privileged groups. A value of 0 indicates equality of opportunity.

Between Group Generalised Entropy Error: Checks whether the algorithm is more fair towards one group compared to another. This metric's population invariance is a desired attribute since vulnerable groups are often also underrepresented in datasets/platforms. This metric is especially sensitive to large discrepancies in the rate of between-groups **positive predictions**. This metric measures fairness not only at a group level but also at the individual level. A value that is larger than 0.01 violates this criterion.

Treatment Equality: Difference in the ratio of **false negative predictions** to **false positives** between two groups. This metric assesses whether, even if the accuracy across groups is the same, is it the case that errors are more harmful to one group than another.

2.3.2 Privacy Preserving Machine Learning

Differential privacy offers a quantitative way to assess privacy protection in algorithms. It allows us to design machine learning models that can be trained responsibly on sensitive data. By incorporating DP, we gain measurable guarantees of privacy, reducing the risk of exposing individual data points within the training set. Roughly, an algorithm is differentially private if an observer seeing its output cannot tell whether a particular individual's information was used in the computation. Differential privacy is often discussed in the context of identifying individuals whose information may be in a database. Although it does not directly refer to identification and reidentification attacks, differentially private algorithms probably resist such attacks (Dwork et al., 2016). In essence, a model trained with DP should be resilient to changes in any single data point or small group of points. This safeguards sensitive information from being revealed during machine learning.

In the presented PPML tool, the efficiency of the DP training procedure is tested using a Membership Inference Threshold Attack, that aims to predict whether a data sample was present in the training data of a machine learning model (Mattern et al., 2023). The success of such an attack is usually measured with 3 different metrics, namely the **Area Under Curve (AUC)**, the **Positive Predictive Value (PPV)** that give the probability that an attacker correctly identifies a random example as well as its correct class assignment (i.e. whether a text excerpt was used for training and whether or not said text was labelled as toxic or non-toxic) and the **Attacker Advantage** that provides a measure about how much additional information the attacker gains with each additional iteration by calculating the difference between the probability of the adversary correctly guessing a data point was included in the training set and the probability of the adversary correctly guessing a data point was not included in the training set (Yeom et al., 2018).

2.3.3 AI Cybersecurity

To rigorously evaluate tools for scanning machine and deep learning models, it's important to specify metrics that objectively measure effectiveness and performance. Classification metrics are crucial; such as **Accuracy**, **Recall**, and **F1-score**, from these **Recall** emerges as a vital metric, especially given the nature of the ITHACA platform and the security focus. Recall measures the frequency with which the tool incorrectly classified actual risky operations as safe. Additionally, employing a **Cost Matrix** can further refine the evaluation by quantifying the impact of different types of classification errors. For example, a cost matrix in this scenario might assign a higher penalty to false negatives (failing to detect risky operations) than to false positives (incorrectly flagging safe operations as dangerous), reflecting the greater potential harm of missing a threat, versus the inconvenience for a human to check a false alarm. These metrics collectively assess the reliability of scanning tools in detecting potential security threats within the models.

Performance metrics are also relevant, such as scan time and resource utilization. The average time it takes to complete a scan provides insight into the tool's efficiency. Monitoring CPU, and memory usage during scans helps evaluate the tool's impact on system resources.

Choosing ModelScan (Protect AI, 2023) as the preferred tool for scanning machine and deep learning models can be attributed to its distinct advantages. Notably, it supports formats like H5, Pickle (Van Rossum, 2020), and SavedModel, and is compatible with major frameworks including TensorFlow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), Keras (Chollet, 2015), Scikit-learn (Pedregosa et al., 2011), and XGBoost (Chen and Guestrin, 2016). This extensive compatibility ensures that it can be integrated into diverse development environments, making it highly versatile for different use cases. One of the critical advantages of ModelScan is its simplicity of operation. The tool can be executed with a single command line, which simplifies its integration

into automated pipelines, enhancing the efficiency of the development process. This ease of use is critical for encouraging consistent security practices across all stages of model development and deployment.

These characteristics collectively make ModelScan a highly effective tool for organizations aiming to enhance the security of their machine learning operations without sacrificing efficiency.

3 Evaluation Tools

One of the main objectives of T5.1 is the implementation of three scalable evaluation tools that would render more fair, private, and secure AI-based systems associated with the ITHACA platform. In deliverable D5.3, the finalised version of the (i) AI Fairness tool for evaluating the impartiality of an AI model against people from marginalised and vulnerable groups, (ii) Privacy Preserving Machine Learning tool for maintaining the confidentiality of sensitive data that would be processed from AI-based systems and (iii) AI Cybersecurity tool for the detection of security breaches and potential risks in AI platforms, are provided. Furthermore, the finalized version of a Visual Component for human moderators that enables the depiction and supervision of the actions and events produced by the functionality of the three aforementioned tools, is developed, being in line with human oversight principle as mentioned in Section 2.2.1.2. **Along with the Moderator's Interface, an extension of the Visual Component was designed for the end users of the platform, to provide them with feedback on the evaluation tools' operational status with adherence to the principles of fairness, privacy and security as defined in Section 2.2. This information will be updated on a real timebasis, so that the users will have ongoing access to the status of the evaluation tools.**

For the purposes of the development of the AI Fairness and PPML tools, a Natural Language Processing (NLP) model for toxicity detection in short written excerpts as the ones the users will post on the ITHACA platform, was created. The toxicity sensor was particularly selected since it was deemed one of the most desirable features for the ITHACA application during citizen jury', municipality's, city managers' and administrators', as well as AI experts' workshops as analyzed in D2.1 and mentioned in Section 1.2 of this deliverable. As for the AI Cybersecurity tool, it's important to note that to our knowledge, no dataset specifically designed for scanning models exists. This poses a significant challenge in evaluating the performance of scanning tools. To address this gap, we create some dummy models that incorporate risky operators to simulate security scenarios. These dummy models lack real-world functionality but deliberately incorporate risky operators, such as unauthorized file access or transmission, that simulate security vulnerabilities found in malicious code.

Furthermore, from D1.3 and specifically T1.2, where classification and analysis of PPML as well as tools that can be adopted alongside AI platforms to preserve their legal compliance, and trustworthiness and manage risk and governance issues, were made, we extracted open-source tools that were utilized as a base for the implementation of the three tools (e.g. Tensorflow Privacy (Google, 2019) as a PPML mechanism). Also, knowledge from T1.2 regarding the nature of the PPML tool (differential privacy over anonymity, federated learning, multi-party computation, and homomorphic encryption) and T1.4 in D1.3, was exploited to choose the metrics for the AI Fairness tool.

3.1 AI Fairness Tool

3.1.1 Evaluation of an AI Toxicity Detection Model using the AI Fairness Tool

The finalised version of the AI Fairness tool employing the metrics described in 2.3.1 was developed and tested upon a Natural Language Processing model for toxicity detection, meaning a mechanism for detecting toxic, discriminatory, threatening and harassing pieces of sentences in a text. As its name indicates, an NLP model captures the meaning of a short written excerpt and reaches a conclusion about the excerpt (Pass, 2023), in our case that is whether a post is toxic or non-toxic. The proposed NLP model was developed in Python using the Keras framework, which significantly speeds up the development of Machine Learning models, by providing easy-to-use and well-documented functions (Chollet, 2015). The newly developed model is a feed-forward deep

learning network that has a text vectorization layer, which transforms a batch of natural language strings into a representation understandable by the model, an embedding layer, which converts each word into a dense real-valued vector of fixed length, and the values of such vectors correspond to the semantic meaning or relationship between the words, a Global Average Pooling layer to calculate the average output of each feature of the model and a Dense (fully-connected) layer with a sigmoid activation function. The model was trained using the Adam (adaptive moment) optimizer, since it provides a faster convergence, it is efficient as well as robust and not memory intensive (Kingma, 2014). The NLP model was trained using the Toxic Tweets Dataset (Iyer, 2021), which contains over 54,000 tweets labelled as toxic (24,153 - 44.8%) or non-toxic (32,592 - 60.36%) resulting in a well balanced dataset. The dataset was synthetically enhanced with the “vulnerability” feature via a random process guided by the results of the citizen workshops conducted in WP2, which indicated that about 20% of the population could be considered as belonging to a vulnerable group, this feature represents whether a tweet belongs to a user that comes from a marginalised or vulnerable group. The latter feature was added to the dataset in order to test whether the model infers a correlation between toxic tweets and authors from vulnerable groups and, thus, simulates potential discriminatory behaviour. It is important to note that since the vulnerability feature did not exist on the original dataset, the process by which it was added was fully randomised. This operation was necessary to test the developed model using the AI Fairness metrics described in Sec. 2.3.1, however in the datasets that will be collected following the development of the platform, this feature will not be artificial, since this information will be available for each user/citizen.

For training the model 80% of the dataset was used, while the remaining 20% was employed for testing the final prediction accuracy of the NLP neural network. Special care was taken to avoid potential cross contamination between the training and testing dataset splits, which would erroneously skew the results. The prediction accuracy of the model is approximately 96%, which indicates that there is a low probability for the model to classify falsely a post and, thus, false negatives or false positives would not affect gradually its fairness.

The metrics of *disparate impact*, *statistical parity difference*, *equal opportunity difference*, *between group generalised entropy error*, and *treatment equality* to evaluate whether our NLP model indicates fairness, were utilised. All metrics were implemented in Python language and were aggregated into one Fairness tool, which has a generic character and it is applicable to any AI-based system. For each metric, the default threshold suggested by the AIF360 library, as well as the authors of the respective metrics, where applicable, as described in D1.3 was set; 0.8 for disparate impact, 0.1 for statistical parity difference and equal opportunity difference, and 0.01 for between group generalised entropy error and treatment equality, as described in Section 2.3.1. If the model achieves a value above this threshold, then it is considered unbiased. For instance, if equal opportunity difference is not achieved, then potential disparities in equal access to positive outcomes (i.e. classifying a post as non-toxic) are suggested. As for the results, due to our enhanced biased dataset, the model reflected the disparities and was rendered unfair by the metrics, indicating the validity of the tool.

3.1.2 Explainability component

Binary classification of text at the paragraph level can aid the ITHACA platform users in deciding whether to remove a specific paragraph or establish guidelines for handling toxic comments. However, for moderators, identifying the offending sections within lengthy paragraphs can be a time-intensive task. For example, if a community member posts a long comment that includes both constructive content and offensive language, binary classification can pinpoint the offending paragraphs, allowing moderators to focus their attention on addressing the harmful content while leaving the rest untouched. This approach not only saves valuable time but also reduces the risk of overlooking toxic comments buried within extensive text.

Moreover, by automating this process, clearer guidelines can be established on what constitutes harmful content, ensuring more consistent and objective decision-making. This in turn reinforces community standards, allowing moderators to act swiftly and fairly when enforcing rules. Additionally, with less manual intervention required, the moderators can focus on more complex or nuanced issues, improving the overall health and atmosphere of the community.

Classifying an entire text excerpt as toxic or non-toxic however, often fails to provide moderators with insight into the exact reasons for the classification. To address this, we developed an explainable toxicity detection system. After identifying a text as toxic, the system highlights or underlines specific toxic words or phrases, pinpointing the elements responsible for the classification (Ribeiro, 2016).

Our solution is designed to streamline automated text moderation, promoting healthy and inclusive communication by reducing the effort required to locate toxic content. It empowers platform moderators to quickly identify and evaluate problematic portions of comments to determine whether the text should be approved or rejected. Additionally, this approach enables more granular toxicity analyses by examining patterns in toxic excerpts, which can inform strategies for mitigating harmful behavior.

The detection of toxic spans in this context is facilitated via a post-processing approach, which is based on a frequency analysis of words that occur in comments classified as toxic, in the training dataset. A list of each different word encountered in the training dataset is constructed, and each word is associated with a toxicity rate i.e. the number of times this specific word was encountered in toxic comments.

The list is constructed after initially pre-processing each comment to remove non-word utterances such as emojis or sighs. Such excerpts may indeed be contextually important in classifying a text piece as toxic, however their interpretation is much more nuanced and as such may not be of much help to the moderator/commenter when highlighted.

3.2 Privacy Preserving Machine Learning Tool

The Privacy Preserving Machine Learning tool developed for the ITHACA platform, employs the notion of DP, so that the data gathered by the users in order to train the highly requested feature of Toxicity Detection ([link/cite study here](#)) is used in a way such that a potential attacker that could gain access to the model, cannot obtain sensitive information about the original users and their respective posts on the platform.

To ensure consistency with the Adam optimizer used for training the fairness-focused NLP model, we employed Differentially Private Adam (DP-Adam). This variant achieves privacy guarantees by incorporating 2 modifications into the standard Adam algorithm. First, it employs gradient clipping to limit the influence of individual data points within a mini batch. This essentially caps the impact any single data point can have on the model's learning direction. Second, DP-Adam injects random noise into the clipped gradients. This noise makes it statistically improbable to determine if a specific data point was included in the training set, simply by comparing how the model updates with and without that particular point. By combining these techniques, DP-Adam allows the model to learn from the overall trends in the data while protecting the privacy of individual data subjects.

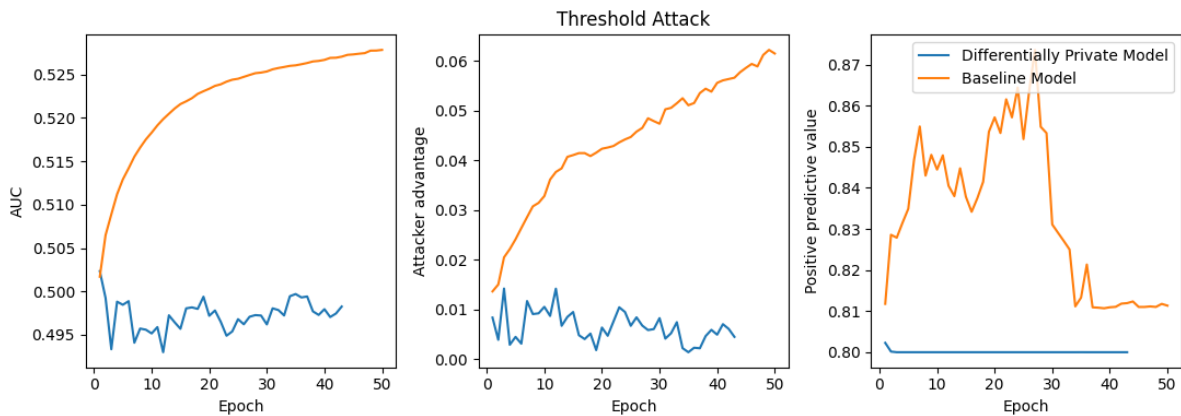


Figure 1: Visualization of the metrics for assessing the level of privacy of the PPML tool, namely, Area Under Curve, Attacker Advantage and Positive Predictive Value.

Our results are shown in Figure 1, where using the DP-Adam optimizer improves all three metrics for the threshold membership inference attack, confirming the applicability of the PPML tool developed for training models developed for and using data by the ITHACA platform, guaranteeing increased privacy for the users of the platform, without requiring additional effort by the users themselves.

The disadvantage of using this method is the potential decrease in the accuracy of the model, however by careful tuning of the hyperparameters of the training procedure only a small decrease of 12% was observed. With the baseline model achieving an accuracy of 96% and the differentially private model achieving an accuracy of 78%, which is considered acceptable for the purposes of the task of detecting potentially toxic user posts on the platform.

3.3 AI Cybersecurity Tool

ModelScan (Protect AI, 2023) is proposed as the AI Cybersecurity tool to be used with the ITHACA platform. It is an open-source tool that scans machine learning models to identify any unsafe code, supporting multiple model formats including HDF5 and SavedModel for Tensorflow, Pickle for Pytoch and Scikit-learn and JSON for XGBoost. This tool is critical for safely using models from frameworks like PyTorch, TensorFlow, Keras, Scikit-learn, and XGBoost. Unlike typical loading procedures that might inadvertently execute malicious code, ModelScan safely checks models by reading file contents byte-by-byte, similar to reading a string. This method allows it to identify unsafe code signatures without triggering them, providing a rapid scanning capability that completes in seconds—the time it takes to read the file from disk. ModelScan categories identified risks into four levels of severity: CRITICAL, HIGH, MEDIUM, and LOW (as shown in Figure 2). Although some code may be embedded in models to facilitate reproducibility for data scientists, this can introduce security vulnerabilities. Users must assess whether embedding code in models aligns with their security standards and operational needs.

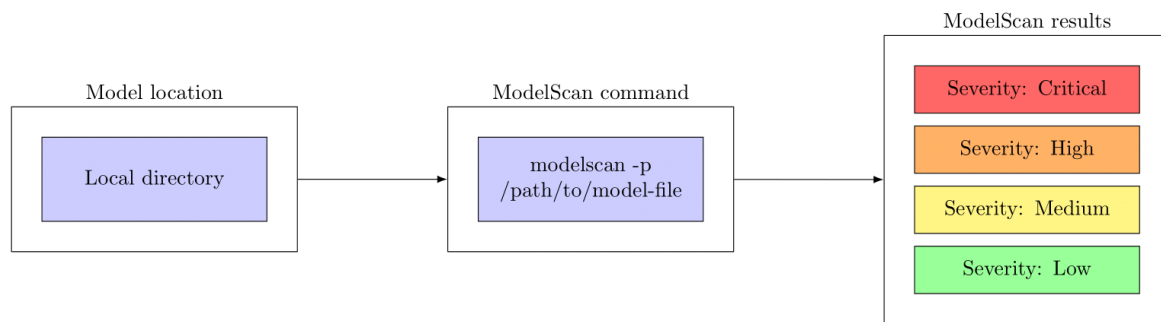


Figure 2: An example of operation of ModelScan.

To highlight the importance of scanning ML models, a TensorFlow model was created that tries to access files within the system. ModelScan performs a thorough analysis by scanning various components of the saved TensorFlow model. Then, ModelScan identifies the use of the unsafe `ReadFile` operator from the TensorFlow module and highlights it as a HIGH severity issue because it poses a significant security risk. Those scans ensure that models do not compromise security or violate data protection. By embedding ModelScan into the development and deployment pipelines, organizations can better secure their AI applications against potential exploits that could leverage deserialization vulnerabilities.

However, while ad-hoc scanning of models when first downloaded is a commendable practice, it alone is not adequate for ensuring robust security within production Machine Learning Operations (MLOps) processes. It is crucial to instill the habit of performing these security checks regularly. Continuous vigilance is necessary because the security landscape and attack vectors evolve rapidly.

To ensure comprehensive security, model scanning must be an iterative process throughout the lifecycle of a model:

1. **Pre-Training Scans:** Scan all pre-trained models before they are loaded for further development work. This step helps to prevent a compromised model from impacting your modeling environment or data science workflows.
2. **Post-Training Scans:** After a model has been trained, it's essential to scan it again to detect any supply chain attacks that could compromise the model during its development.
3. **Pre-Deployment Scans:** Before deploying a model to production, a final scan should be conducted to ensure that no compromises have occurred during storage or any other phase post-training.

These steps form a critical part of securing machine learning operations and protecting against RCE. Figure 3 shows a visual representation of the recommended scanning pipeline for machine learning models, designed to emphasize the critical checkpoints where ModelScan should be applied.

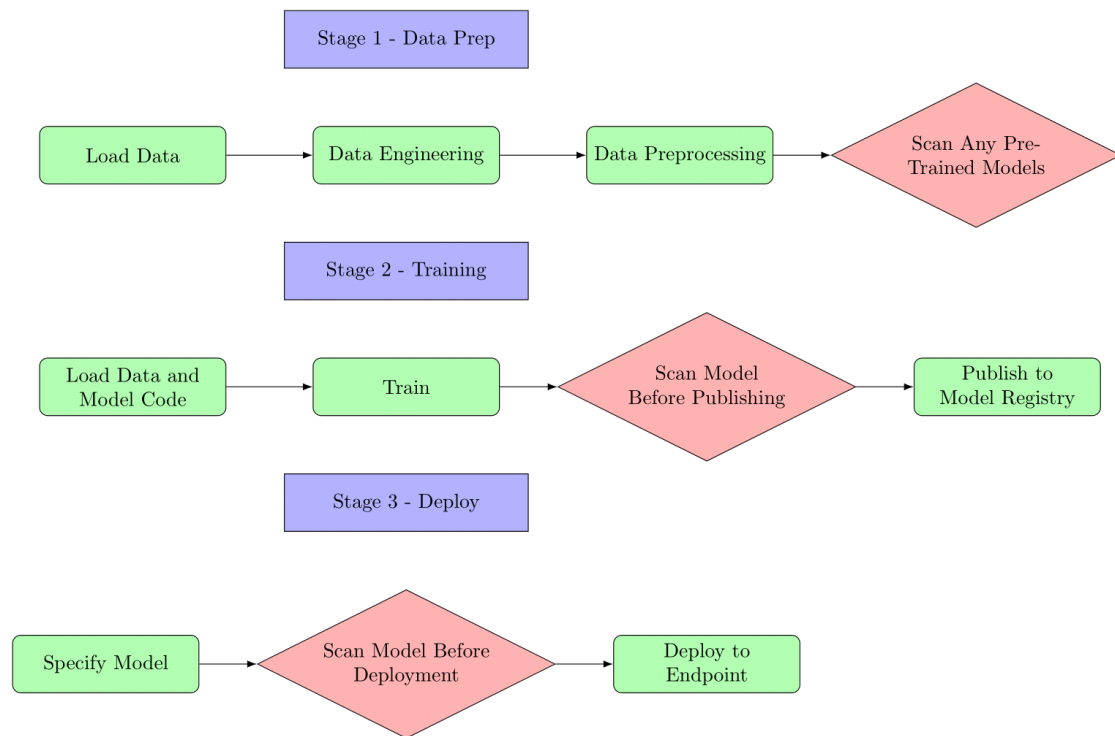


Figure 3: The steps of the scanning pipeline for machine learning models.

To align with the ITHACA platform requirements, we have developed a script that integrates ModelScan into machine learning tools, such as the toxicity detection tool. This script automates the scanning of models for security vulnerabilities, ensuring that only safe models are loaded and saved within the system. The script performs the following actions:

- 1. Preload Scan:** Before loading a model, the script analyzes the model file for any unsafe code or security issues. If the model is deemed safe, it proceeds to load the model using the specified framework (TensorFlow or PyTorch). If not, it alerts the user (which can either be the moderator, notified via the moderators' interface, or a citizen notified through the respective visual component) and prevents the model from being loaded, thus safeguarding the system from potential threats.
- 2. Pre-Save Scan:** When saving a model, the script first saves a temporary version of the model and scans it for security risks. If the temporary model passes the scan, it proceeds to save the model permanently. If the scan identifies any issues, the script deletes the temporary model and notifies the user, preventing unsafe models from being saved and distributed.
- 3. Compatibility:** The script supports both TensorFlow and PyTorch frameworks, making it versatile for different deep learning frameworks. It can easily be integrated into various development environments that use these frameworks.
- 4. Verbose & Logging:** Users have the option to enable verbose output, which provides detailed information about the scanning process and any issues detected.

3.4 Visual Component

3.4.1 Moderator Interface

For the Moderator Visual Component, an interface was developed that can be used by a responsible party (i.e. human moderator) to access and monitor the required tools and, thus, have control over big data and events arising from the three aforementioned tools.

This final version of this interface provides the following functionalities:

- Highlighting words heavily associated with toxic posts by the AI Fairness tool with a color from a palette from green to red indicating their level of toxicity. This feature will aid the moderator in monitoring the functionality of the AI Fairness tool, while highlighting text as an indication of toxic speech detection complies with qualitative criteria “C” in Section 2.2.1.2, regarding legal obligations on **human oversight and transparency** in AI mechanisms and decisions (Art. 13 par. 3 (iv) and Art. 14 of the AI Act). **This mechanism was modified to further enhance explainability as described in Section 3.1 in Deliverable D5.4.**
- Visualizing metrics such as *Attacker Advantage* to evaluate the level of privacy that the PPML tool offers, while the authors of posts depicted in the visual component remain anonymous to maintain impartiality, fairness, and privacy. **This feature was also updated with outlining the best values for these privacy metrics and their respective plots are shown with the press of a button (“PPML Metrics” button as shown in Figure 4).**
- Visualizing the output of the AI Cybersecurity tool by providing a comprehensive list of security indications as outlined in Section 3.3 (classifying detected security vulnerabilities and threats in ITHACA AI systems as CRITICAL, HIGH, MEDIUM, or LOW).
- Visualizing tools’ status information such as the date of their last update, operation, security and privacy status. The last two are marked with a Green/Orange/Red indication providing feedback to the moderator regarding the state of the tool under inspection (Fully Secure/Private, Some Security/Privacy Criteria Are Not Fulfilled, Not Secure/Private, respectively). This feature offers a holistic overview of the functionality of the three evaluation tools, AI Fairness, PPML and AI Cybersecurity tool, hence facilitating the human moderation process, which is crucial as defined in Art. 14 of AI Act.

In Figure 4 a prototype version of the visual component is illustrated. A thorough example of how a moderator may utilize this component as well as a description of the features added to this final version is outlined in deliverable D5.4.

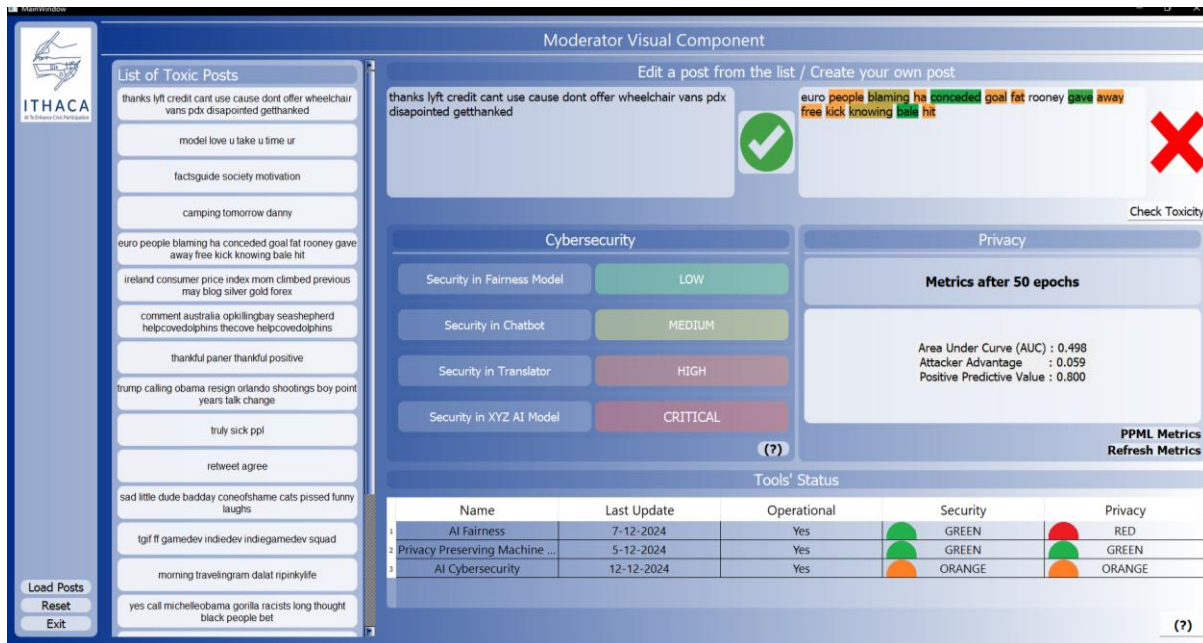


Figure 4: The prototype version of the Moderator's Visual Component.

The final version of the visual component aids the moderator with the supervision and maintenance of the underlying tools, in the context of human oversight in AI models. Nevertheless, the outcomes of the evaluation tool can be incorporated into the UI of the platform to provide the citizens with feedback regarding the AI's decision-making processes and generally the functionality of the three evaluation tools in a clear, understandable and accessible manner, according to Art. 13 par. 3 (iv) of the AI Act, “Transparency and provision of information to deployers”. Thus, the tools developed for task 5.1 are available to be used as an API during the development of the platform, to both evaluate the underlying AI models that the ITHACA platform would incorporate. The **explainability/transparency** requirement is fulfilled by the word highlighting functionality provided as an API to be integrated in the platform.

3.4.2 User Interface

To further adhere to the legal requirements of **transparency and explainability** as described in Section 2.2.1.2, an auxiliary User Interface to provide the users with comprehensive visual notifications and descriptions of the evaluation tools' outputs and state, is implemented. As depicted in Figure 5, this user-centric visual component gives information to the user concerning:

- How fair, private and secure are the AI models running in the background of the ITHACA platform (in the form of statistics), thereby enhancing the accountability of the respective tools (a principle intertwined with fairness according to Art. 5 par. 1 (a) of GDPR), AI systems and overall platform. As an example, the user will have access to information regarding how impartial and unbiased the Toxicity Speech Detection tool or chatbot, etc., deployed by the ITHACA platform.
- Tools' status. This feature is identical with the one integrated in Visual Component's Moderator Interface.
- Platform statistics arising from the average fairness, privacy and cybersecurity stats.

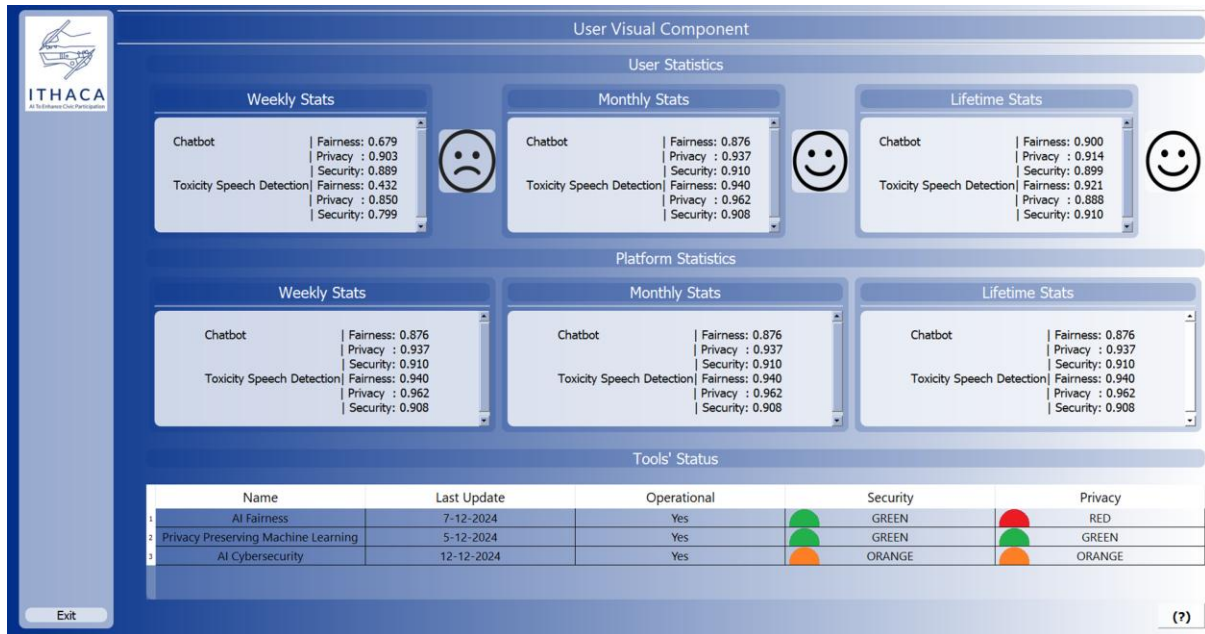


Figure 5: The prototype version of the User’s Visual Component.

This UI may be used as an extension of the Visual Component to inform the user regarding the hidden procedures performed on the backend of the ITHACA platform to support its compliance with the principles of fairness, privacy and security as well as providing an insight to the internal processes of AI systems exploited by the platform. This feature not only builds trust among users and the platform functionalities but also ensures a level of transparency that is essential for enabling users to object or seek clarifications from human moderators. This, in turn, reinforces the principles of fairness and accountability in AI applications. The functionality of this UI component and its features are described in deliverable D5.4 in detail.

3.4.3 Accessibility Features

ITHACA platform would support various accessibility features that are essential to facilitate the experience of users with special needs in compliance with legislations concerning Accessibility and clarity of AI moderation Logic (Art. 13 par. 3 (iv) of AI Act as outlined in Section 2.2.1.1.2 of this document). Features such as:

- Voice navigation and activation as well as providing keyboard navigation to incorporate an inclusive and accessible by all users way to browse the ITHACA platform. For example, the platform would integrate a screen reader that would convey all information included in both Visual Components (the one addressing the moderator and one for the end user) to meet the needs of people with visual impairments.
- Modifications in platform contrast options and colors via the accessibility button for people with visual impairments. Ensure a minimum color contrast ratio of 4.5:1 between standard text and its background, and 3:1 for larger text. Additionally, avoid using color as the sole means of conveying information, in accordance with paragraphs “1.4.1 Use of Color - Level A” and “1.4.3 Contrast (Minimum) - Level AA” of Web Content Accessibility Guidelines (WCAG) version 2.1¹.

¹ https://www.w3.org/WAI/WCAG22/quickref/?versions=2.1¤tsidebar=%23col_customize&levels=aaa

- An accessibility large on screen menu for platform navigation and buttons for both users with special needs as well as elders.
- Alternate text in images would be useful for diverse users to have information to understand what each indication means.

will be incorporated into the ITHACA platform.

These accessibility features will extend to all platform functionalities and UI elements, including Moderator's and User's Visual Component, which will obtain these features upon integration. Apart from the above features, this version of User's & Moderator's Visual Component integrates some accessibility elements such as increased font size (in comparison to the previous version) and text accompanying color indications to aid people with color blindness. Moreover, we added "help" buttons (indicated with the symbol "(?)" as shown in Figure 4 and Figure 5), upon pressing which a pop-up window appears containing simple explanations of the visual indications in the form of phrases that would be comprehensive to all users. Overall, all these features will facilitate the usage of the ITHACA platform by both people with disabilities and, also people belonging to other vulnerable groups, such as elders.

4 Conclusion

In the context of the activities of T5.1 the final version of two supportive tools that are applied alongside AI-based civic participation systems to evaluate the latter in terms of Fairness and Privacy, are developed, while an open-source tool to aid with the conformity of the Security principle in such platforms is proposed. The technical (functional) requirements of these tools were defined after identifying legal (non-functional) requirements specified by EU legislations concerning the employment of AI models in civic engagement systems. Apart from these tools, a Visual Component that acts as a demonstrator of the functionality of the implemented tools, gives the ability for human intervention and provides a level of explainability over the outcomes of these tools as described by the Qualitative criteria, is also implemented.

In this second phase of task T5.1, the main developmental changes to render the evaluation tools in a prototype state, are outlined below:

- Enhanced the explainability mechanism paired with the AI Fairness tool, to render its decision-making process transparent to both the human moderator and the end-user, being consistent with Art. 5 par. 1 (a) (concerning Fairness and Transparency principles) of GDPR and Art. 13 par. 3 (iv) of the AI Act.
- Implemented an API to facilitate the integration of the Cybersecurity tool into AI systems of the ITHACA platform, such as the toxicity detection tool. This wrapper automates the scanning of models for security breaches and threats, thereby ensuring that only secure models are loaded and stored within the platform, being compliant with all legal requirements described in Section 2.2.3.2.
- Populated Moderator's Visual Component with more functionalities to showcase the capabilities of the evaluation tools, while information about the tools' status are also depicted to ease the overseeing procedure.
- Extended Visual Component with a User Interface, to give comprehensive feedback to the users w.r.t. the evaluation tools' compliance to the fairness, privacy and cybersecurity principles and as well as provide information about the status of the tools always respecting the qualitative criteria outlined in as described in Section 2.2.1.2.

As for the linkage of the evaluation tools and the ITHACA platform, these tools concern part of the tools provided for the purpose to meet the technical requirements of the platform (T3.1), and they should meet the needs of the users and the implementation of the project Use Cases (T2.4). At this point it is worth noting that this prototype version of both the AI Fairness tool and PPML developed in the final phase of T5.1 are tested upon a toxicity detection mechanism, which consist one of the most essential features in a civic participation systems as derived by the results of T2.4. Moreover, based on the work carried out in T3.1, the tools are intended to address issues related to the security of both the operation of the platform and the users' interaction with it. More particularly, the technical requirement of security, stipulates that to address the current security issues through the implementation of the Use Cases', it is necessary to address functionalities which should perform:

- Check for spam or "toxic" content.
- Inappropriate content identification
- Guard overall security of the platform and users personal data protection against external sources

The AI Fairness tool is linked to the toxic identification content with the scope to safeguard all content posted across the various fields of ITHACA platform. PPML makes sure that no personal data are identified and/or shared when other AI processes take place in the platform. Hence, this tool is linked to the inappropriate content identification, but also the safeguarding of the user's personal data on the platform when these are treated by automated algorithmic processes. The latter two required functionalities of the technical requirement of security

are assisted by the AI Cybersecurity tool, which monitors and warns against malicious “external” attacks which may compromise not only the breach of personal data, but also the safety-critical operation of the other two tools.

Consequently, we provide the three evaluation tools for conformity with the fairness, privacy and cybersecurity principles in ITHACA AI systems, which tools are conceptualized and implemented taking into account both legal requirements and technical requirements reflecting the users’ opinions and desires. In this final phase of task T5.1, the second and finalised version of the above tools is made available with enhancements addressing deficiencies of the previous version of the tools and modifications to further comply with the qualitative criteria thoroughly presented in Section 2.2. Moreover, the Moderator’s and User’s Visual Component after being integrated into the ITHACA platform would incorporate all platform’s accessibility elements (as mentioned in Section 3.4.3) in an accessible, usable and inclusive interfaces for all citizens. Further details regarding the development and operation of AI Fairness, PPML and AI Cybersecurity tool are incorporated into the updated version of D5.2, namely D5.4.

5 References

A

Abadi, Martin, Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... others. (2016). Tensorflow: A system for large-scale machine learning. In *12th Symposium on Operating Systems Design and Implementation (16)* (pp. 265–283).

Alaya, B., Laouamer, L., & Msilini, N. (2020). Homomorphic encryption systems statement: Trends and challenges. *Computer Science Review*, 36, 100235.

Arachchige, P. C. M., Bertok, P., Khalil, I., Liu, D., Camtepe, S., & Atiquzzaman, M. (2019). Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7), 5827-5842.

B

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671-732. <https://doi.org/10.15779/Z38BG31>

C

Chan, L., Morgan, I., Simon, H., Fares Alshabanat, Ober, D., Gentry, J., Cao, R. (2019). Survey of AI in Cybersecurity for Information Technology Management. <https://doi.org/10.1109/temscon.2019.8813605>

Charter of Fundamental Rights of the European Union, 2000/C 364/01. (2000). *Official Journal of the European Communities*. Retrieved from https://www.europarl.europa.eu/charter/pdf/text_en.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785>

Chollet, F. (2015). Keras.[online] Available at: <https://github.com/fchollet/keras>. Accessed, 14(05), 2023.

D

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2016). Calibrating Noise to Sensitivity in Private Data Analysis. *Journal of Privacy and Confidentiality*, 7(3), Article 3. <https://doi.org/10.29012/jpc.v7i3.405>

E

European Commission. (2020). White Paper on Artificial Intelligence - A European approach to excellence and trust. Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Data Protection Board. (2020). Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. Retrieved from https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf

European Parliament and Council of the European Union. (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 (General Data Protection Regulation - GDPR)*. Official Journal of the European Union. Retrieved from <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>

European Parliament and Council of the European Union. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial

Intelligence Act) and amending certain Union legislative acts. Official Journal of the European Union. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>

G

GDPR - General Data Protection Regulation. (2016). Official Journal of the European Union.

Gillespie, T. (2020). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Google. (2019) TensorFlow Privacy. [Online]. Available at: https://www.tensorflow.org/responsible_ai/privacy/guide. Accessed, 14(05), 2024.

H

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., ... & Zwitter, A. (2019). Will democracy survive big data and artificial intelligence? In *Towards Digital Enlightenment* (pp. 73-98). Springer, Cham.

I

Information Commissioner's Office. (2020). Explaining decisions made with AI. Retrieved from <https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/>

Iyer, A. (2021). Toxic Tweets Dataset. [Online]. Available at: <https://www.kaggle.com/datasets/ashwiniyer176/toxic-tweets-dataset/data>. Accessed, 22(04), 2024.

K

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Nair, V., & Hinton, G. (2010). Cifar-10 (canadian institute for advanced research). URL <http://www.cs.toronto.edu/kriz/cifar.html>, 5(4), 1.

Kullmann, M., & Cefaliello, A. (2021). The Role of Human Oversight in Artificial Intelligence: The Case of Predictive Policing. Retrieved from https://www.law.ox.ac.uk/sites/files/oxlaw/the_role_of_human_oversight_in_artificial_intelligence_0.pdf

L

LeCun, Y., Cortes, C., & Burges, C. (2010). MNIST handwritten digit database.

Liu, Y., Lee, W. C., Tao, G., Ma, S., Aafer, Y., & Zhang, X. (2019, November). Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1265-1282). <https://doi.org/10.1145/3319535.3363216>

M

Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., & Berg-Kirkpatrick, T. (2023). *Membership Inference Attacks against Language Models via Neighbourhood Comparison* (arXiv:2305.18462). arXiv. <https://doi.org/10.48550/arXiv.2305.18462>

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

P

Paaß, G., & Giesselbach, S. (2023). *Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media* (p. 436). Springer Nature.

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32 (pp. 8024–8035). Curran Associates, Inc. Retrieved from <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf> scikit-learn: machine learning in Python. Retrieved June 12, 2024, from Scikit-learn.org website: <https://scikit-learn.org/stable>
- Pedregosa, F., Varoquaux, Gael, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.
- Phan, N., Wu, X., Hu, H., & Dou, D. (2017, November). Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *2017 IEEE international conference on data mining (ICDM)* (pp. 385-394). IEEE.
- Pinto, S. (2021). AI and Human Oversight: How to Make It Work. Retrieved from <https://www.forbes.com/sites/forbestechcouncil/2021/08/25/ai-and-human-oversight-how-to-make-it-work/?sh=6e7f12cd5167>
- Protect AI. (2023) ModelScan: Protection against Model Serialization Attacks. [Online]. Available at: <https://github.com/protectai/modelscan>. Accessed, 18(06), 2024.

R

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

S

- Sarker, I. H., Hasan Furhad, & Raza Nowrozy. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Computer Science/SN Computer Science*, 2(3). <https://doi.org/10.1007/s42979-021-00557-0>
- Schneier, B. (2018). *Click Here to Kill Everybody: Security and Survival in a Hyper-connected World*. W.W. Norton & Company.
- Sunstein, C. R. (2001). *Republic.com 2.0*. Princeton University Press.

V

Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation.

Vassilev A., Oprea A., Fordyce A., & Anderson H. (2024). NIST AI 100-2 E2023, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. <https://doi.org/10.6028/nist.ai.100-2e2023>.

W

Woolley, S. C., & Howard, P. N. (2016). Automation, algorithms, and politics| political communication, computational propaganda, and autonomous agents—Introduction. *International Journal of Communication*, 10, 9. <https://ijoc.org/index.php/ijoc/article/view/6294>

X

Xu, R., Baracaldo, N., & Joshi, J. (2021). Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint arXiv:2108.04417*.

Y

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on Large Language Model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 100211–100211. <https://doi.org/10.1016/j.hcc.2024.100211>

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282. <https://doi.org/10.1109/CSF.2018.00027>

Z

Zapechnikov, S. (2020). Privacy-Preserving Machine Learning as a Tool for Secure Personalized Information Services. *Procedia Computer Science*, 169, 393-399. <https://doi.org/10.1016/j.procs.2020.02.235>