Grant Agreement No.:

823783
Call: H2020-FETPROACT-2018-2020
Topic: H2020-FETPROACT-2018-01
Type of action: RIA

D9.4 A Report Identifying Possible Abuses of WeNet ("Misuse Scenarios Report")

Revision: v.0.3

Work package	WP 9
Task	Task 9.4
Due date	31/12/2022
Submission date	21/12/2022
Deliverable lead	Jessica Heesen
Version	0.3
Authors	Laura Schelenz, Jessica Heesen
Reviewers	Amalia de Götzen, Loizos Michael

Abstract	This deliverable discusses five misuse scenarios that could arise in the context of operating the WeNet platform and further opening it to external developers and researchers. We discuss the following cases of abuse: hate speech, a lack of privacy and data protection, scientific misconduct using WeNet data, discriminatory designs of apps and services, and child sexual harassment or the distribution of child pornography. We focus on the protection of WeNet end users who are registered in the WeNet Hub and use

	WeNet apps. The deliverable also provides recommendations on how to prevent abuses of WeNet as well as draft codes of conduct for researchers and designers.
Keywords	Ethics, platform, abuse, misuse scenario, content moderation, scientific integrity, non-discrimination, recommendations

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	30/11/2022	1st draft	Laura Schelenz
V0.2	12/12/2022	Final draft from WP9	Laura Schelenz, Jessica Heesen
V0.3	12/19/2022	Reviews incorporated	Laura Schelenz, Amalia de Götzen, Loizos Michael

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the "WeNet - The Internet of US" (WeNet) project's consortium under EC grant agreement 823783 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	**R**	
Dissemination Level		
PU	Public, fully open, e.g. web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to WeNet project and Commission Services	

** R: Document, report (excluding the periodic and final reports)*

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

CONTENT

Co-funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

Co-funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Co-funded by the Horizon 2020
Framework Programme of the European Union

EXECUTIVE SUMMARY

The WeNet platform is a **public space for learning, social interaction, research, and innovation**. It is open to communities in Europe and around the world. The use of the WeNet platform for the WeNet pilots (see deliverables 7.1 and 7.2) and Open Call projects (see deliverable 8.4) show that the platform can be a **valuable tool to connect researchers, developers, and end users** from diverse cultural contexts to improve services for university students and the wider population. In a public space such as the WeNet platform, which welcomes participation of existing contacts of the WeNet consortium as well as external researchers and developers, abuse can occur. We **define abuse as any act of attack, exploitation, deception, and exclusion of end users** of the WeNet platform, which thereby violates the values, trust, and integrity of the WeNet community.

Abuse can take many forms. In this report, **we focus on five misuse scenarios**:

1) Abusive communication including hate speech in the WeNet apps
2) Lack of transparency about privacy rights and data collection practices
3) Scientific misconduct including fabrication, falsification, and plagiarism
4) Discrimination through design and de-facto exclusion of WeNet end users
5) Abuse of children (every person under the age of 18 according to the UNCRC) including cybergrooming and child pornography

We have deliberately excluded scenarios related to hacking, cybersecurity, and data theft. For considerations on these topics, we refer to our prior work that has extensively dealt with privacy and data protection concerns (see deliverable 9.3). In the five misuse scenarios mentioned above, the group of persons subjected to the abuse are WeNet end user. While the WeNet community consists of three types of users (1. researchers, 2. developers, and 3. end users), we believe that **WeNet end users are the most vulnerable to attacks, exploitation, deception, and exclusion**. They are not in a position of power compared to the influence of researchers and designers on the WeNet community. The responsibility to protect WeNet end users from abuse can be derived from European human rights and data protection frameworks.

Based on the discussion of these scenarios, **we provide recommendations for the prevention of abuse**. We anticipate that the WeNet platform will grow in the number of apps and services offered, and that the WeNet community will expand to include researchers and developers who are external to the WeNet consortium and Open Call winners. In anticipation of this growth, it may be prudent to implement the following safeguards against abuse in the WeNet platform.

First, a **sound concept for content moderation** should be established. This is a weak spot in global platform management and requires the exploration of creative methods including community self-regulation. Second, WeNet should **implement transparency via annual transparency reports and mechanisms in the WeNet apps** that help users understand how data is collected and used to optimize services for the end user. Third, the WeNet platform should **require researchers and designers leveraging the infrastructure of the WeNet platform to sign a code of conduct** that is geared towards scientific integrity and non-discriminatory design, respectively. Fourth and finally, WeNet should **establish roles for the handling of (allegations of) abuse**.

These roles can be filled by one or more persons, and they may need to collaborate with law enforcement.

These recommendations are the four high-level recommendations, further fleshed out in chapter 7, and **each recommendation can be implemented in different ways**. Therefore, throughout the deliverable, there are suggestions on how to realize these recommendations. They serve to inspire the WeNet consortium but are not exclusive to considerations of preventing abuse.

While the research and thoughts laid out in this deliverable may be relevant for other social media platforms, they have not been designed with the broader social media industry in mind. Instead, they **consider the framework that shapes WeNet: the European research and innovation context**. This context demands that results from the WeNet project are open to the public and that the WeNet platform is as inclusive and participation-oriented as possible. This context differs from other social media tools that may be governed by private companies and corporate interest. In this sense, lessons learned from this deliverable can inspire the broader social media industry but recommendations are specific to the WeNet context.

Co-funded by the Horizon 2020
Framework Programme of the European Union

1. INTRODUCTION

This section sets the stage for this deliverable. We define the focus of our approach to preventing abuse of WeNet. The primary subjects of our considerations are WeNet end users, i.e. those who use, test, and engage with the WeNet apps built on top of the WeNet platform. They are also the people who may use the WeNet open online course for educational purposes as part of a class with their professor. WeNet end users may both be affected by abuse from researchers, developers, and other end users, and be abusers when it comes to sharing inappropriate content or hate speech. The deliverable will consider the complexity of this situation in exploring ways to prevent abuse. For the sake of this deliverable, we understand abuse as any act of attack, exploitation, deception, and exclusion of end users of the WeNet platform, which thereby violates the values, trust, and integrity of the WeNet community. The responsibility to prevent abuse stems from the European human rights framework including the Charter on Fundamental Rights, as well as European frameworks for the responsible design and development of Artificial Intelligence (AI).

1. THE WENET PLATFORM AND STAKEHOLDERS

The WeNet Platform brings together developers, researchers, and students around the opportunity to create and test new applications, learn about computer science, design, and innovation, and collect or utilize data for research on societally relevant topics. The WeNet platform, which can be accessed via the WeNet Hub, hosts several apps. They are developed by computer scientists or student developers of the WeNet community (to date consisting of members of the WeNet consortium and winners of the Open Call competition). The end users are students in the different university pilots, and they are testing new applications. Additional end users from the broader population may be users brought in through the Open Call competition. Researchers can also use the WeNet platform by accessing data in the WeNet Research Infrastructure. The data collected via the different WeNet apps is anonymized and freely accessible to registered researchers, who may want to conduct scientific analyses with the data.

Based on this configuration of the WeNet community, we have three groups of “users” of the WeNet platform: developers, end users, and researchers (cf. D 7.2, “User Recruitment Procedures”, page 6). **This deliverable focusing on misuse scenarios is concerned primarily with the protection of end users from possible abuses of WeNet**. End users are the most vulnerable group participating in the WeNet community because they are the ones subjected to the designs of the different apps in the WeNet platform and their data is collected for research and development purposes. They are also on the front lines of interacting with other end users in the WeNet apps. Talking about potential abuse, **it is then end users who must be protected from potential abuses by peer end users as well as by developers and researchers**. This affirms the multifacetedness of abuse and protection in an online platform. While end users themselves can be abusers, they also need to be protected from abuse. Furthermore, given the position of power that developers and researchers find themselves in vis à vis end users, the protection of end users must be a priority to ensure a safe and welcoming environment for the general public.

Co-funded by the Horizon 2020 Framework Programme of the European Union

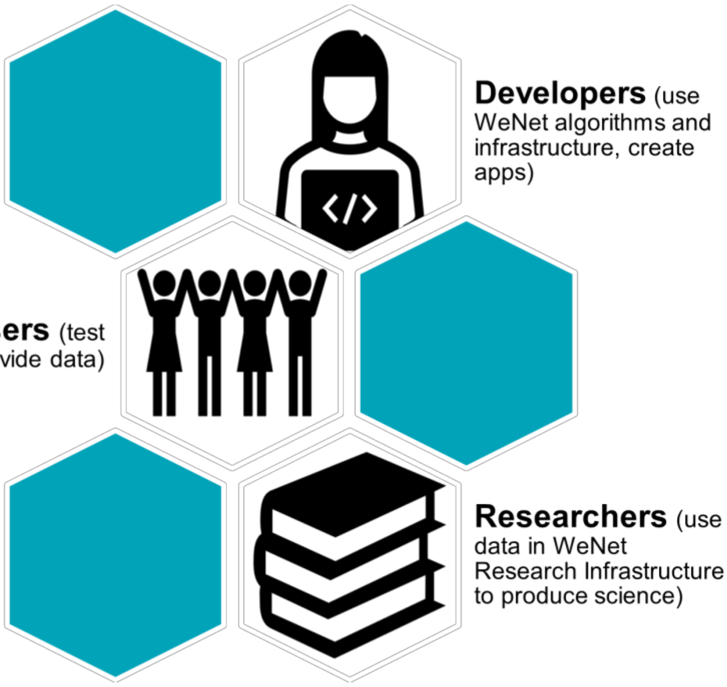

Developers (use WeNet algorithms and infrastructure, create apps)
End Users (test apps, provide data)
Researchers (use data in WeNet Research Infrastructure to produce science)

FIGURE 1: USERS OF THE WENET PLATFORM

2. A WORKING DEFINITION OF ABUSE

What constitutes an “abuse” of WeNet? This question requires contemplation on the meaning of abuse in the context of social media platforms. Abuse can refer to a variety of unethical and illegal behaviors on the Internet and in the context of creating technological solutions for the public: “Definitions of abuse fast become embroiled in contentious debates around privacy, freedom of speech, democracy, discrimination and the power of big tech companies” (Vidgen et al., 2019, p. 10). However, in the literature on abuse in social media platforms, abuse is understood to be more specifically about speech. “Abuse” frequently refers to an inappropriate or criminal act of digital communication or interaction, including online sexual harassment and sexual attacks, identity theft, stalking, tracking, recording, revenge porn, and trolling (Lumsden & Harmer, 2019, 4ff). In addition, abuses can fall under the category of hate speech. They constitute gender-based attacks, attacks against LGBTQ+, or anti-Black, anti-Muslim, and antisemitic violence (Brown, 2021). Experts differentiate between interpersonal abuse (personal attack on an individual) and group-directed abuse (defamation of a social identity, usually a marginalized or under-represented group in society) (Vidgen & Derczynski, 2020, p. 6). In the context of hate speech, abuses are also differentiated according to their gravity to avoid a binary construction of abuse and non-abuse. For instance, a classification could distinguish abusive content from offensive content from non-abusive content (Vidgen & Derczynski, 2020, p. 7).

Following a narrower understanding of abuse, **we may define abuse of WeNet as “any act of violent digital communication or interaction directed at an individual or a group of peer end users.”** This implies that the violation is executed by another end user or a group of end users in one or several of the WeNet apps. As laid out in

Co-funded by the Horizon 2020 Framework Programme of the European Union

chapter 1.1, we focus on the protection of end users. This should involve the protection of end users from other end users. **Content moderation is thus a critical question and will be picked up in chapter 2 of the deliverable**.

However, abuse can include more scenarios than speech or communication and interaction between end users. Abuse can also relate to the **(non-)provision of privacy guidelines and the (disproportionate) collection of data**. In such a case, the trust of end users is abused as is the right to privacy and data protection of the end user. Another form of abuse is the implementation of **discriminatory technology**, whether in the WeNet profile or in the independent WeNet apps. Discrimination may occur due to bias in the underlying datasets used to train a system or bias in design features. One example is the development of an application that excludes impaired end users because it is not compatible with screen reader software used by blind users. This case can be considered an abuse if the design team leverages the WeNet platform for the advancement of its own technology with no regard for the European values and ethical guidelines associated with WeNet. Here, again, trust of the WeNet community is abused and power is exerted for individual gain. Finally, researchers may collect and use data in the WeNet platform to conduct questionable analyses that are done sloppy at best. Scientific integrity and the responsible use of data for research may be at risk when third-party researchers gain access to WeNet data at a large scale. It is then important to consider safeguards against the **misuse of data from WeNet end users for 'bad science.'**

Based on the additional considerations of possible abuses above, we extend our definition of "abuse" to a broader version: **"An abuse of WeNet constitutes any act of attack, exploitation, deception, and exclusion of end users of the WeNet platform, which thereby violates the values, trust, and integrity of the WeNet community."** The deliverable will shed light on different scenarios of such abusive acts and provide guidelines for preventing abuse.

3. ETHICAL RESPONSIBILITY TO PREVENT ABUSES OF WENET

The WeNet project is committed to complying with the following ethical standards:

1) EU Charter on Fundamental Rights
2) European Convention for the Protection of Human Rights and Fundamental Freedoms
3) European Commission (n/a) Ethics, Horizon 2020 – The EU Framework Programme for Research and Innovation, https://ec.europa.eu/programmes/horizon2020/en/h2020-section/ethics
4) European Commission (2007) Ethics for Researchers. Facilitating Research Excellence in FP7. Luxembourg: Publications Office of the European Union.
5) Helsinki Declaration
6) UN Convention on the Rights of the Child (UNCRC)
7) UNESCO Universal Declaration on Bioethics and Human Rights
8) Nuremberg Code

These standards lay out core human rights and civil rights protections that require the prevention of abuses in the WeNet platform.

The freedoms laid out in the Charter on Fundamental Rights (right to liberty, security, respect for private life, protection of personal data, freedom of expression, and freedom of assembly) demand that **end users are protected from abuses in the WeNet platform that infringe on their physical or mental integrity**. Moreover, the dignity of WeNet end users (Art. 1 of the Charter) but also freedom of speech must be protected, which is why the WeNet platform requires a sound concept for content moderation.

The EU document on *Ethics for Researchers* lays out the responsible use of data from research subjects. The guideline includes information on informed consent, ethics review procedures, and "research involving developing countries." This document is relevant and provides a baseline for **preventing abuses by researchers of data gathered from WeNet end users** through the various apps in the WeNet platform. All research leveraging the data generated in and through WeNet must comply with ethical standards and principles for excellence in research.

A particular responsibility is the **protection of minors per the UNCRC**. Art. 3 of the UNCRC states, "In all actions concerning children, whether undertaken by public or private social welfare institutions, courts of law, administrative authorities or legislative bodies, the best interests of the child shall be a primary consideration." Children are persons under the age of 18 as defined by the UNCRC. Minors may be active in the WeNet platform because student populations in different pilot sites (different countries, also outside the European Union) may be under the age of 18 but still participate in the WeNet platform. Future apps may also be designed specifically for children given the shift to digital learning since the Covid 19 pandemic. This **requires WeNet to take actions in the best interest of the child**. We therefore consider a distinct misuse scenario concerning the group of children. Chapter 6 elaborates on how to ensure that minors do not become the victims of inappropriate contact, child grooming, and child pornography.

In addition to these general responsibilities to prevent abuse of WeNet, there are more specific ethical standards that relate to the context of computer science, technology development, data science, automation, and artificial intelligence.

1. EU General Data Protection Regulation EU 2016/679
2. Communication on a comprehensive approach on personal data protection in the European Union [COM(2010) 609]
3. The directive on privacy and electronic communications (58/2002/EC)
4. Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data
5. Ethics Guidelines for Trustworthy AI, HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE
6. Communication on Artificial Intelligence for Europe, COM(2018) 237 final

The Communication on AI for Europe stresses the **importance of trust and accountability when developing AI**. This means that entities like the WeNet platform which rely on artificial intelligence in their innovations must provide mechanisms that

Co-funded by the Horizon 2020 Framework Programme of the European Union

build trust, including mechanisms to hold the designers responsible and to account. Earning the trust of the public is vital for engagement with the platform. Preventing abuse is a central factor in building and keeping trust of WeNet community members, especially end users.

Privacy and data protection play a crucial role in building and upholding trust as well. The GDPR provides clear rules for the collection, storage, and processing of data from end users. Not simply from an ethical but from a legal perspective, the adherence to **privacy by design and default** is required. Abuses of such principles can amount to legal challenges including fees and sanctions. Preventing abuse of WeNet end users' data is thus key for ethical but also economic and practical reasons.

The Ethics Guidelines for Trustworthy AI reiterates the fundamental freedoms in the Charter on Fundamental Rights but in addition identifies **four core principles that can be derived from the Charter in the context of artificial intelligence**: Respect for human autonomy (no manipulation, no coercion or deception and always keeping a human in the loop), prevention of harm (mental and physical integrity of users), fairness (no unfair discrimination or bias, equal opportunity), and explicability (transparent communication, audits).

The Ethics Guidelines further stress **privacy and data protection as core ingredients to the development of trustworthy AI**. On page 17 of the report, it says: "Digital records of human behavior may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them." Privacy then relates not only to strictly legal requirements of informed consent, secure storage, and retention periods, but also to the **ethical use of data**. This is an important aspect in the context of the WeNet platform. The diversity of applications and the amount of data gathered (on personal aspects of life such as daily routines, food and alcohol intake, academic performance, etc.), renders WeNet end users vulnerable to re-identification.

On page 17 of the report, it is further stated that "the quality of the data sets used is paramount to the performance of AI systems." Preventing abuse can thus also be related to **ensuring a good quality of data**, ensuring that labels are correct, missing data is marked or false data is eliminated from the dataset before it is used to train algorithms or conduct research.

In summary, the European human rights framework and the European frameworks for responsible design and development of AI offer the rationale for preventing abuse of WeNet:

- Upholding the **dignity** of WeNet end users is key
- Ensuring the **mental and physical integrity** of WeNet end users is key
- Protecting **fundamental freedoms** including life, liberty, and privacy of WeNet end users is key
- Complying with legal **privacy and data protection** rights of WeNet end users is key
- Protecting **children** among WeNet end users is a priority
- Ensuring **responsible and excellent research** with WeNet data is key

- Ensuring good **quality of the data** produced in and through WeNet is key

4. THE FIVE MISUSE SCENARIOS

In the following, five misuse scenarios are presented. They were created based on prior discussions with WeNet colleagues at various consortium meetings. They are **realistic but have not yet occurred** to the best of our knowledge. They are scenarios that are likely to take shape in a larger, significantly grown platform with more end users, researchers and developers that are not closely connected to the core WeNet consortium.

The scenarios are not the only ones possible to imagine. However, they cover **many situations where the trust of WeNet end users in the WeNet community is negatively affected**. Some scenarios are entangled with legal questions. While legal requirements will be considered, they cannot be debated conclusively because the authors of the deliverable are not lawyers but ethicists.

The scenarios include:

1) Abusive communication including hate speech in the WeNet apps

In this scenario, a WeNet end user decides to disengage from the WeNet platform because they see abusive language in one of the messages received from another end user.

2) Lack of transparency about privacy rights and data collection practices

In this scenario, a WeNet end user finds out that their data is fed into the research infrastructure. They were not aware of this because information about data collection was not transparent enough for this end user. They loose trust in the WeNet community.

3) Scientific misconduct including fabrication, falsification, and plagiarism

In this scenario, a WeNet end user reads about the production of falsified scientific articles with WeNet data in the news. They are disappointed that their data is abused in this way and leave the WeNet platform.

4) Discrimination through design and de-facto exclusion of WeNet end users

In this scenario, a blind user experienced discrimination because an app developed in the WeNet platform was not optimized for screen reader software. Designers had not considered the needs of imaired users which caused harm.

5) Abuse of children (every person under the age of 18 according to the UNCRC) including cybergrooming and child pornography

In this scenario, an underage WeNet end user connected to an adult interested in exploiting children. Since the adult had created a fake account and acted as a child, the underage WeNet end user exposed to them was in danger of cybergrooming and sexual abuse.

Co-funded by the Horizon 2020 Framework Programme of the European Union

The following chapters will explore ways to prevent these scenarios and ultimately lead to four high-level recommendations for the WeNet platform.

Co-funded by the Horizon 2020 Framework Programme of the European Union

2. MISUSE SCENARIO: ABUSIVE COMMUNICATION

Scenario 1: Misuse of WeNet's opportunities for open exchange

WeNet end user A engages in one of the apps offered as part of the WeNet platform. The app is based on the exchange of academic recommendations for reading material in smaller groups of interests. WeNet end user B posts a recommendation for the reading of an antisemitic book and adds a hateful comment to this post. WeNet end user A sees the post and is disturbed and offended by the abusive communication. They don't know what to do and immediately leave the WeNet platform.

A well-recognized problem in online social media platforms is the dissemination of hate speech by users of the platform in posts, messages, shares, and comments (Daniels, 2013; Marantz, 2019). Hate speech has the potential of ruining an online community. Moderation of content is thus key to ensure that users are protected from abusive peers. "Left unchallenged, abusive content risks harming those who are targeted, toxifying public discourse, exacerbating social tensions and could lead to the exclusion of some groups from public spaces" (Vidgen & Derczynski, 2020, p. 1). Hence, "the health of online communities can be severely affected by abusive language" (Vidgen & Derczynski, 2020, p. 5).

While the detection and removal of abusive content is key for healthy online communities, the regulation of abusive content demands constant attention to real-time on-going communication in the online community. "Editorial work takes time" and the operators of a platform may experience enormous pressure to fulfil expectations of regulation (Heesen, 2021). This means that preventing abusive communication in WeNet-related products requires an around-the-clock supervision of the communication in apps developed on top of the WeNet platform. Providing this oversight through the employment of humans may be costly. Automation is a common approach taken by platform managers. The benefits and limitations of an automated approach to content moderation (possibly based on natural language processing models) will be discussed below.

A clarification is necessary here. There are strict legal requirements for large social media companies to remove abusive content within 24 hours (Vidgen & Derczynski, 2020, p. 5). These legal requirements may, however, not be applicable to the WeNet platform because of its lower scale. Nevertheless, the WeNet community aspires to meet ethical principles that require self-commitment and self-regulation. Especially in a service developed with public funds, ethical principles such as equal access and equity in participation should be met. Hence, the following arguments and points to consider are provided from an ethical perspective, not a legal one. Additionally, there are ethical questions about where to draw the line when it comes to freedom of expression. While these questions cannot be discussed in detail in this deliverable, future research should tackle the meaning of freedom of speech in an increasingly connected (as in 'using online services') but politically polarized European society.

1. CHALLENGES IN THE AUTOMATED RECOGNITION OF ABUSIVE COMMUNICATION

There is an array of literature and datasets on abusive language, including hate speech, racist speech, and sexist speech (Gibert et al., 2018; Gorrell et al., 2018; Vidgen & Derczynski, 2020). Computational tools help detect abusive language in social media platforms and flag the respective content or remove the content altogether. Abusive language is recognized by a scan for pre-defined keywords. This often happens partially or fully automated with little or no involvement of humans.

There are several challenges to the automated recognition of abusive language. Previous studies have raised concerns about bias and discrimination (European Union Agency for Fundamental Rights, 2022, p. 50). In the context of this debate, one case is benevolent speech that carries discriminatory potential. Jha and Mamidi (2017) show that Twitter data contains a significant amount of derogatory expressions about women, yet does not constitute abusive language per se (pp. 8ff). These expressions, which fall in the category of benevolent sexism, read like a positive post, a compliment even, but carry harmful gender stereotypes (Jha & Mamidi, 2017, p. 7, also see table 1). Since these expressions do not contain harmful terms, the system may not be able to detect and flag them as problematic (Jha & Mamidi, 2017, p. 11).

Sexist assumption	**Benevolent Expression**	**Hostile Expression**
Women are not as intelligent as men.	She's amazing, handling that senior position like her male colleagues.	I can't understand why she holds that position, she's a dumb [abusive word].
	Not flagged!	**Flagged!**

TABLE 1: EXAMPLES OF BENEVOLENT AND HOSTILE SEXIST SPEECH (CF. JHA & MAMIDI, 2017)

In addition, Jha and Mamidi (2017) found in their research that, "since benevolent sexism seems harmless, noble, and even romantic at times, it is retweeted more number of times as compared with tweets that exhibit hostile sexism" (10). This leads to another concern: If a certain message containing benevolent sexism (which is thus not flagged but potentially offensive to users) is shared widely, this message may even be picked up and further distributed by machine learning algorithms due to its popularity.

While benevolent speech is primarily used here to exemplify the structural limitations of AI-based (hate) speech recognition algorithms, it is an interesting case for consideration under a WeNet content moderation concept. Benevolent stereotyping may not fall under the category of abusive communication per se, given the strong protection of freedom of speech in liberal democracies. However, it can contribute to toxifying an online community. WeNet should therefore discuss different types of speech (and decide how to deal with them) when developing a sound concept of content moderation.[1]

[1] Another point to consider is the WeNet capacity to conduct scientific experiments that help us understand how many users are using benevolent vs. hostile expressions. One of the central advantages of WeNet is its potential to act as a mini-world for empirical studies in social settings. This

A strong limitation to automated hate speech detection is that existing computational tools may in effect silence marginalized or minority communities. Davidson et al. (2019) found that algorithms trained to detect and flag harmful content disproportionately detected sexism and racism in African-American speech. This may be the case because language frequently used by Black people contains words that the training data linked to negative expressions (Davidson et al., 2019, p. 30). Additionally, Black-aligned language uses expressions such as the n-word in an empowering sense, whereas the n-word is rightly banned in mainstream speech (Rahman, 2012). The algorithmic system may not be able to distinguish between speech that uses the n-word in an empowering or derogating sense. This goes to the difficulty of AI to understand context. Hence, one of the major challenges is that the automated recognition and handling of abusive communication lacks attention to context. If hate speech catalogues are applied without context, then we have to expect them to disproportionately flag minority communities such as African American people (Davidson et al., 2019, p. 32).

Request with abusive intent	Request with empowering intent
Looking for peers to start a band, no [word] please.	Looking for fellow [word] to start a social justice project on campus.
Flagged	**Flagged**

TABLE 2: CONTEXT MATTERS IN THE DETECTION OF HATE SPEECH

In table 2, we see two examples of speech that would be recognized by a computer system as containing abusive language. The first statement is undoubtedly offensive: a derogatory term is used against a social identity (group-directed abuse, cf.Vidgen & Derczynski, 2020, p. 6) and used to exclude people. This content would rightly be flagged and removed by the system. The second statement is a call for other Black people to join a campaign against racist violence on campus. The flagged term is used in an empowering sense and refers to an in-group social identity. This content would wrongly be flagged and removed, constituting a form of censorship. In effect, this would further oppress a social group that wanted to use the social media tool for a good cause.

Irony is another difficult matter. Irony, sarcasm, and humor is context-dependent, plus highly subjective. Detection of irony in verbal speech may be difficult, let alone in written communication and by a computer (Vidgen & Derczynski, 2020, p. 14).

Given these challenges of the automated recognition of abusive communication, WeNet may want to consider alternative or complementary measures that mitigate the risks described above. At the very least, it seems appropriate to have a human-

opportunity could be leveraged in understanding the motivation of end users and the need for content moderation.

Co-funded by the Horizon 2020 Framework Programme of the European Union

machine ensemble that works together – the machine suggests flagging and removal, and the human understands context and decides.

2. SOCIAL MEASURES MITIGATING THE RISK OF ABUSIVE COMMUNICATION

With the limitations of automated detection of abusive communication, we should **consider alternatives that are grounded in social practices of community management**. The goal is to foster an atmosphere of free exchange in the WeNet platform and apps, while reminding end users that freedom demands responsibility. The community-based approach proposes social measures to curb hateful comments and disrespect in the WeNet platform. First, end users engaging in WeNet must understand that the system does not automatically scan for abusive language. Disclosing this information is relevant to the end users' evaluation of risk when engaging with the apps in the WeNet platform. Second, the WeNet community should be empowered to regulate itself. This means that WeNet end users should hold each other accountable for abusive language. In such a scenario, end users could produce counter-narratives that call out the abuser and debunk their content as abusive. Third, to ensure that the WeNet platform remains a space where the law is obeyed, it may be **necessary to employ humans in the loop who check communication in the WeNet platform and apps for criminal activity** such as sexualized violence and hate speech. Depending on the number of apps offered in the WeNet platform, these measures may require a combination of automated and social measures.

1. Transparency and complaint mechanisms

Transparency can help end users understand how the system works and what to expect when engaging with it. "Transparency is a practice of system design that centers on the disclosure of information to users, whereas this information should be understandable to the respective user and provide insights about the system. Specifically, the information disclosed should enable the user to understand why and how the system may produce or why and how it has produced a certain outcome (e.g. why a user received a certain personalized recommendation)" (Schelenz et al., 2020). **Transparency is thus geared towards explaining the system and helping the end user understand how they are affected by it**.

In the context of the WeNet platform, transparency measures can be taken to inform the user about the context of communication in the WeNet platform. For instance, before a WeNet end user starts engaging with an app in the WeNet platform, **the system could send out a 'warning' that such communication is not subject to automated content moderation and has not been checked for abusive language**. The disclaimer could read: **"Messages sent in this app are forwarded to you as they are. They are not checked for abusive language."** Such information helps WeNet end users 'brace themselves' and understand that the system has not deliberately amplified abusive content.

In addition to information about the mode of the message, the user needs to have control over the message. Following Schelenz et al. (2020), user control is "the possibility of users to interact with the system to adjust elements thereof to their respective needs and preferences" (p. 24). One solution that respects user control in the WeNet platform and apps is to allow the end user to report an abusive message. The system must thus provide an option for a complaint/report mechanism.

"We just forwarded this request to you as is. It was not checked for abusive language. Did you find this message offensive?"

"We just forwarded this request to you as is. Please let us know in case you found this request offensive." + button labelled **"Report"**

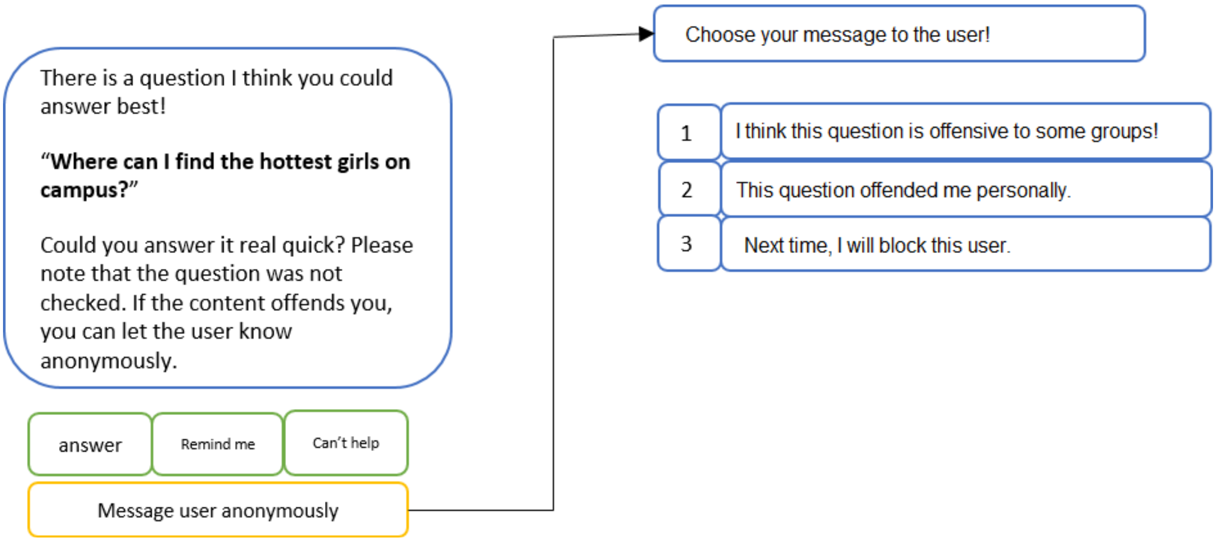
2. Counter-narratives and community self-regulation

Once WeNet end users are informed about the context of communication, and potentially reported an offensive message, the question becomes: **who should deal with the abusive communication and how?** In a community-based approach, members of the community self-regulate their speech. For instance, someone from the community takes the role of the person responsible for regulation. These positions can be voluntary and alternate between community members. "The participants of public online communication are […] nodes of a network that are mutually responsible for each other and owe each other communicative demands such as truthfulness and clarity" (Heesen, 2021, p. 440). **End users of apps in the WeNet platform may thus hold each other accountable in a relational manner**.

One possibility to **call out abusive communication is to use counter-narratives**, i.e. stories or statements that counter the hate in abusive communication. Previous research has suggested automated counter-narratives in online social media platforms to counter hate 'from below.' Chung et al. (2019) compile a multilingual dataset of responses to hate speech. These may include factual arguments that disrupt the hateful narrative. They cite literature that asserts this method of fighting hate speech is the most effective (p. 2819). So far, the method has been used by trained human operators, mostly NGO workers. Automating counter-narratives in large-scale environments may be a faster and more effective way to reach communities in online social media platforms (Chung et al., 2019, p. 2820), but the biases mentioned with regard to automated hate speech detection (inability to flag benevolent offenses, lack of context-awareness) persist because the system still has to identify abusive content.

A more appropriate way to implement counter-narratives may thus be via community self-regulation. For instance, the recipient of an abusive message in a WeNet app may have the option to **anonymously respond to the author of the abusive content and point out that they feel offended by the content**. These messages may be pre-formulated and the end user may choose among them.

Co-funded by the Horizon 2020 Framework Programme of the European Union



There is a question I think you could answer best!

“Where can I find the hottest girls on campus?”

Could you answer it real quick? Please note that the question was not checked. If the content offends you, you can let the user know anonymously.

Choose your message to the user!

1 I think this question is offensive to some groups!

2 This question offended me personally.

3 Next time, I will block this user.

FIGURE 2: EXAMPLE OF A CHATBOT FLOW (IN THE WENET APP “ASK FOR HELP”) FOR RAISING AWARENESS ABOUT HATEFUL CONTENT THROUGH A COUNTER MESSAGE

3. Accountability and ensuring safety and security under the law

Finally, although the WeNet community is best advised to take responsibility for its own healthy interaction, the WeNet platform must provide a lawful framework to do so. This means that criminal activity such as sexualized violence, hate speech, or – as discussed in chapter 6 – child grooming, must be curbed. Depending on the scale of the WeNet platform and the number of users in the WeNet apps, a combination of automated detection of illegal content and human oversight may be warranted. If the use of a computational tool to detect, flag and/or remove abusive content is envisioned, **this tool should be built on training data that comes from the WeNet community**. The pilots involved in the WeNet project involve countries across the world as well as different cultural backgrounds and languages. To apply e.g., a hate speech filter, the taxonomy that defines hate speech must be informed by all pilots involved. Best practices for creating a dataset on hate speech can be found in Vidgen and Derczynski (2020, p. 18).

Humans should always be in the loop to review cases of abusive communication and make decisions about keeping an end user who breaks the rules in the platform. Community counter-narratives from peers constitute a soft form of accountability. Ultimately, though, the system operators must decide whether they ban an abusive end user from the platform. Again, banning a user based merely on a computational identification of abusive language may accidentally lead to the exclusion of minority or marginalized communities (Davidson et al., 2019). Criteria for the banning of users and human oversight are required to make a fair decision about exclusion from the online community. Furthermore, if the abusive message constitutes a hate crime, law enforcement must have access to the identity of the user in order to prosecute the case. Given the complexity of regulating speech in a publicly funded and open space, questions remain and should be discussed in several sessions with the WeNet

Co-funded by the Horizon 2020 Framework Programme of the European Union

consortium as well as lawyers before taking further steps to opening the WeNet platform to the broader public.

3. MISUSE SCENARIO: LACK OF TRANSPARENCY

Scenario 2: Lack of Information about users' privacy rights

WeNet end user A signs up to the WeNet platform because they want to use a particular app that was advertised at their university. They agree to the terms and conditions of using the WeNet platform by ticking a box. Once signed up to the WeNet platform, the user again agrees to the terms of using the app. With these two instances of consent, the end user received all information about data collection and use affecting them. However, due to this one-time and obscure presentation of information, end user A misses the fact that their data is fed into the WeNet research infrastructure. When they find out through conversation with another WeNet end user that their data is used to construct a dataset on university life, they are disappointed about the WeNet platform and feel that their privacy rights have been violated.

Privacy and data protection are core rights in the European legal and human rights framework (see Charter on Fundamental Rights and EU GDPR). However, while legal compliance may be ensured, WeNet end users may nevertheless feel abused. This can be the case if privacy policies and data collection activities are not obvious to WeNet end users. A **lack of transparency about WeNet's privacy and data collection practices may cause ignorance among end users about** the collection of their data for different purposes. This **lack of transparency affects end users' autonomy and right to information**. If end users cannot access and understand relevant information about privacy and data collection practices in the WeNet platform, they cannot make an informed and independent decision about whether and how their data should be used.

Previous research has identified deliberate design practices that circumvent transparency as so-called "dark patterns." Wagner et al. (2020) look at German regulation of social media networks. They warn of the use of so-called **dark patterns by companies to render ineffective the transparency requirements** provided by law. Citing Gray et al. (2018), Wagner et al. (2020) understand dark patterns as "instances where designers use their knowledge of human behavior (e.g., psychology) and the desires of end users to implement deceptive functionality that is not in the user's best interest." There is then an intentional character to the obscuring of information, and intentional misleading of end users that can amount to manipulation. Wagner et al. (2020) stress the importance of implementing transparency legislation and persecuting violations, e.g. the fine given by the Germany's Federal Office of Justice to Facebook for their intransparency.

This **intentional misleading of end users must be considered an abuse** which is subject to this chapter. To prevent possible abuse, it should be anticipated that designers leveraging the WeNet platform may deliberately try to mislead end users in order to collect more data from users. A less extreme scenario is that designers leveraging the WeNet platform may be unaware of the importance of transparency and lack guidance about implementing transparency in the WeNet platform. Despite no negative intentions, the results for WeNet end users may be the same, namely that they do not understand their privacy rights and how their data is collected and used in WeNet. To prevent abuses of WeNet end users' privacy and data rights, it is thus important to highlight the importance of transparency in system design and provide guidance to designers (Felzmann et al., 2020).

1. THE IMPORTANCE OF TRANSPARENCY

Let us first recapitulate what transparency means. The literature on transparency and the adjacent field of 'explainable AI' generated a diversity of definitions (Ehsan et al., 2021; Felzmann et al., 2020; Mohseni et al., 2021; Nassih & Berrado, 2020; Tintarev & Masthoff, 2015; Yu & Li, 2022; Zerilli et al., 2022). These definitions vary in their scope of what transparency includes (e.g. is it just about disclosing information or also about its effects on the end user). The above cited authors generally understand **transparency as a requirement or a practice that helps the user understand something about the system**, and providing information is at the heart of all definitions cited above.

In our deliverable, we would also like to highlight the definition of transparency developed by WeNet colleagues in their work Schelenz et al. (2020). The definition (as stated earlier in Chapter 2.2.1) reads: "Transparency is a practice of system design that centers on the disclosure of information to users, whereas this information should be understandable to the respective user and provide insights about the system. Specifically, the information disclosed should enable the user to understand why and how the system may produce or why and how it has produced a certain outcome (e.g. why a user received a certain personalized recommendation)." This definition has been generated from a review of literature in technology ethics and information systems (Schelenz et al., 2020). It combines the 'disclosure of information' (which is dominant in the works of Zerilli et al. (2022) and Tintarev and Masthoff (2015)) with the requirement that this information must be understandable to the user because it is not actionable otherwise (see Yu and Li (2022)).

A concept that is **closely related to transparency is user control**. According to Schelenz et al. (2020), user control refers to "the possibility of users to interact with the system to adjust elements thereof to their respective needs and preferences." The concept of user control is similar to what Tintarev and Masthoff (2015) call "scrutability" or Kulesza et al. (2015) call "controllability." These concepts have in common that they allow users to influence the system in some way. Why do transparency and user control matter?

Transparency and user control are **important for the perceived trustworthiness of an AI-based technology as well as for user satisfaction**. Studies have established a relationship between transparency and user trust (Branley-Bell et al., 2020; Kizilcec, 2016; Molina & Sundar, 2022; Schmidt et al., 2020; Shin et al., 2022; Yu & Li, 2022; Zhao et al., 2019). This relationship is complicated as the 'right' level of trust and the preferred modalities of explaining the system to the user can vary among individual users or groups of users. When implementing transparency, it is important that the information is neither oversimplified nor too complex (Zerilli et al., 2022; Zhao et al., 2019), that it is presented in an appropriate way (e.g. the combination of visual, textual, and spoken elements through an avatar or agent (Mohseni et al., 2021; Weitz et al., 2021)), and adapted to the literacy and understanding of the individual user or a group of users (Felzmann et al., 2019; Guesmi et al., 2021).

The ‘subject’ of transparency and user control can vary. For instance, explaining a decision or outcome of an AI-based process can be crucial in some domains. This is the case in medical decision-making (Angerschmid et al., 2022; Wang et al., 2019) or when it comes to the AI-based recruitment of people for employment (Kong et al., 2021). However, **privacy and data protection rights should always be made available to the end user of a system** or the person subject to the system’s processing (Heesen et al., 2022). Transparency about data collection processes and user control mechanisms that allow users to opt out of the collection of some data are key. The following sections of this chapter go deeper into the ways that WeNet can provide transparency about its privacy policies and data collection practices.

2. INTRODUCING A WENET TRANSPARENCY REPORT

WeNet may consider releasing an annual transparency report to inform its stakeholders about its core practices. Transparency Reports are **publications that inform the public about the practices of a company, an organization or an institution**. They are usually published annually. In the context of the technology industry, transparency reports are released by “internet and mobile ecosystem companies”, telecommunications companies, and companies working on “new technology” (Access Now, 2021). They may include information about data and privacy policies, data management and data centers, inclusion and diversity policies at the company, standards for advertisement, criminal activity in a platform, countermeasures taken by the company to fight illegal content or behavior, and statistics of users, stakeholders, etc.

An **example of a big tech transparency report is the Google transparency report** which can be found at the following link: https://transparencyreport.google.com/ This report compiles information about the handling of user data, including information about third parties such as governments, national security agencies, or companies asking for access to Google user data. The report also includes information about the removal of content due to copyright infringements, data protection violations or hate speech. Furthermore, there is a section on security in Android, safe browsing, Google’s fight against child abuse, political advertisement rules, and interruptions in Google services.

Previous research has investigated users’ expectations towards information about social media platforms in platform transparency reports. Luria (2022) surveyed users specifically on the topic of disclosing recommendation algorithms and how recommendation and personalization affect the user. They find that **users want direct and specific information,** and that **examples, interactive or visual representations of how data is used for personalization help users** understand the system. In a co-designing activity with participants, Luria (2022) develops four interface designs to guide authors of transparency reports. These designs can inspire WeNet platform owners and designers in providing their own annual transparency report.

Inspired by the literature and examples above, a transparency report about the WeNet platform should include the following information:

- **How many active users** are there in each user group of the WeNet platform: A) how many developers of apps? B) how many end users of apps? C) how many researchers using WeNet data?
- **How many apps are active** and can be used in the WeNet platform and who are the designers of these apps?
- What are the 'house rules'/what is the **code of conduct** in the WeNet platform? How are these rules enforced? What happens if a user breaks these rules?
- **How many rule breaks** of the WeNet platform code of conduct were recorded during the year? Was there **criminal activity** in the WeNet platform? Was there cooperation with law enforcement?
- What are the **data protection and privacy policies** in place in the WeNet platform? Who stores the data in the WeNet platform and where is it stored (country)?
- How is the **data of end users used and (how) can end users influence** data collection practices (e.g. opt in/opt out and complaint mechanisms)?

If such information about practices in the WeNet platform was available to the public (and especially WeNet end users), **stakeholders of the WeNet platform could develop more realistic expectations about their interaction in the platform**. Information must be provided in clear and simple language to be accessible to the broader public.

3. TRANSPARENCY *IN* END USERS' INTERACTION WITH THE WENET PLATFORM AND APPS

While annual transparency reports may be an appropriate way to provide comprehensive information in a single 'file' or at a single link, such information has to be actively searched for by interested stakeholders. It is not displayed during the end users' interaction with the platform, where most questions arise and decisions are made about providing data. It is thus **not enough to rely on a 'global' document for transparency in the WeNet platform**. Rather, transparency information should be integrated into the end users' interaction with the WeNet platform (Schelenz et al., 2022).

Previous research has explored **interactive interfaces that help users understand on the spot why they are seeing a certain piece of content** (Kasote & Vijayaraghavan, 2020; Kleanthous et al., 2019; Kulesza et al., 2015; Zheng & Toribio, 2021). In these intelligent user interfaces, the end user can also adjust elements of personalization or recommendation. They can thus provide immediate feedback to the system about how their data is used. Typical instruments that help users understand

and adjust the system's workings to their needs and preferences include filters or control panels and pop-up windows or unsolicited messages.

To offer a **more concrete example** about how transparency measures can be integrated into the interaction of WeNet end users with the WeNet platform and apps, let us explore an example case. This **case concerns the existing WeNet application "Ask for Help," which is a chatbot embedded in Telegram**.[2] The chatbot allows WeNet end users to ask any question into the WeNet community. The technology underlying the WeNet platform then picks a relevant end user from the pool of WeNet end users registered in the app "Ask for Help."[3] This chosen end user's profile matches some desired characteristics that help answer the question posted by the other end user. The question is thus forwarded to the end user who can then decide to answer it. The profiles of WeNet end users are built based on the collection of survey data about the user's routine habits, including their hobbies, their academic performance, their transportation, their socializing, their eating behavior, and more.

Example: An explanation for receiving a question in the WeNet app "Ask for Help"

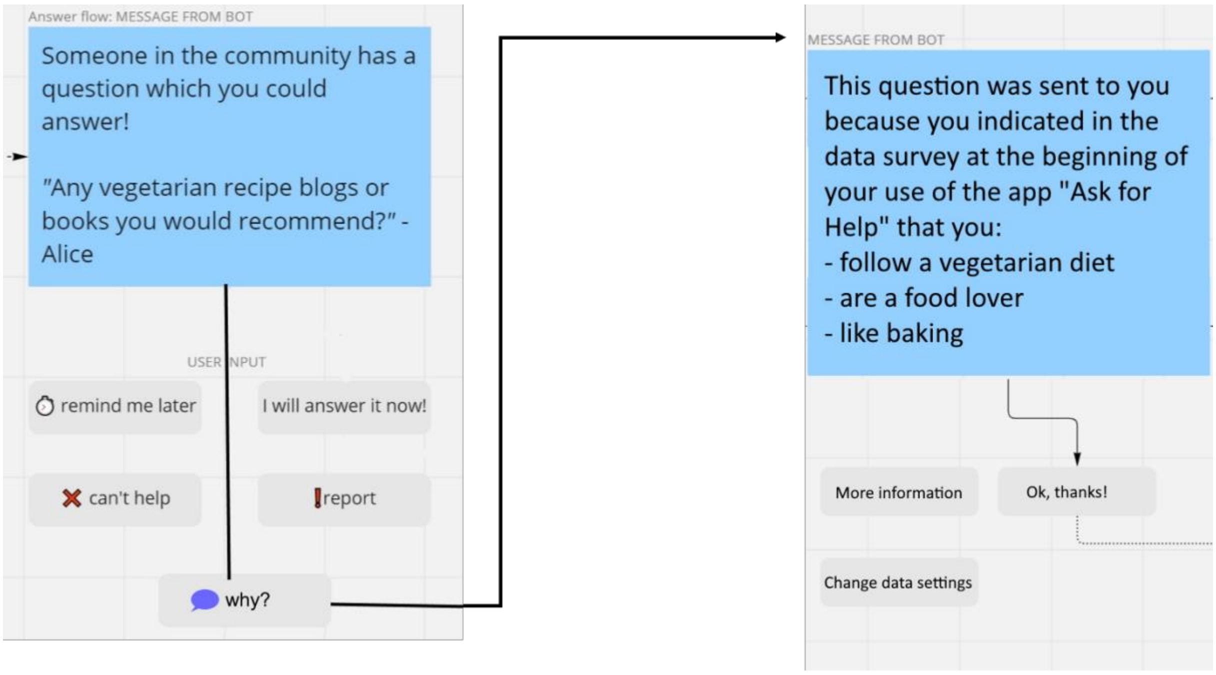
One possibility to implement transparency in the WeNet platform is to provide an **explanation to the WeNet end user why they were forwarded a particular question**. Musto et al. (2019) offer inspiration in this regard. In their research paper, they design explanations in a recommendation bot that offers movie recommendations to users. Just like the WeNet app "Ask for Help," their "MovieRecSysBot" is embedded in Telegram. Users receive recommendations through the Telegram chat, e.g. by the system posting a movie's poster advertisement and adding core information such as length of the movie, director, actors, and genre of the movie. The user then can like or dislike the recommendation, see details about the movie, skip this recommendation, and receive feedback about why they are seeing this recommendation. Recommendations are made based on the users' past ratings. This is made transparent through the design feature "why?", which can be accessed with the click of a button. Musto et al. (2019) offer an example explanation for the recommendation of the movie "The Untouchables": "I suggest 'The Untouchables' because you like films where: the director is Brian de Palma as in Casualties of War; the genre is Thriller as in The Departed; the musicComposer is Ennio Morricone as in Casualties of War" (p. 105).

In the WeNet chatbot "Ask for Help," **a "why?" feature could be a valuable addition to the user flow**. The current chatbot flow offers the following reactions to receiving a question from a peer end user: remind me later; I will answer it now; can't help; and report. It could be expanded by including an option "why" that explains why the user has been chosen as a qualified respondent to this question. **Figure 3 shows an example user flow that integrates transparency**. In this example flow, the end user can act upon the explanation by either demanding more information or changing the

[2] It should be noted that Telegram may lack ethical practices of information disclosure and privacy notification itself. Using the Telegram platform in WeNet apps may thus amplify challenges to the WeNet end user's autonomy and right to information.

[3] This chatbot was called "Ask for Help" only in the earlier versions of the chatbot. Later, the name was adapted to "We@universitypilot," for example "We@LSE." To be consistent and for simplicity, we stick to the earlier name "Ask for Help" throughout this deliverable.

settings that relate to the use of their data for the matching algorithms in “Ask for Help.” The “more information” option will lead the user to general information about WeNet’s data collection practices such as an annual Transparency Report (see 4.2). The “change data settings” option will allow the user to control what kind of data is used by the system to decide which requests are forwarded to this user.



Someone in the community has a question which you could answer!

"Any vegetarian recipe blogs or books you would recommend?" - Alice

This question was sent to you because you indicated in the data survey at the beginning of your use of the app "Ask for Help" that you:
- follow a vegetarian diet
- are a food lover
- like baking

FIGURE 3: EXAMPLE USER FLOW OF AN INTEGRATED EXPLANATION FOR RECEIVING A CERTAIN QUESTION IN THE WENET APPLICATION “ASK FOR HELP”

Co-funded by the Horizon 2020 Framework Programme of the European Union

4. MISUSE SCENARIO: SCIENTIFIC MISCONDUCT

SCENARIO 3: PUBLICATION OF FALSIFIED OR LOW-QUALITY SCIENTIFIC ARTICLES USING WENET DATA

WENET END USER A AGREES TO THE COLLECTION OF THEIR DATA FOR RESEARCH PURPOSES. THEY KNOW ABOUT THE DATA COLLECTION PROCESSES IN THE WENET PLATFORM AND ARE HAPPILY PROVIDING INFORMATION INTO THE WENET RESEARCH INFRASTRUCTURE TO SUPPORT SCIENCE. TWO YEARS AFTER STARTING TO USE AN APP IN THE WENET PLATFORM, THEY READ ABOUT A CASE OF SCIENTIFIC FRAUD WITH WENET DATA IN THE NEWS. APPARENTLY, ACADEMICS HAD USED WENET DATA TO PRODUCE FALSIFIED RESEARCH AND ONE OF THE RESEARCHERS HAD SECURED A POSITION AS A POLITICAL ADVISOR TO GOVERNMENT X. WENET END USER A IS DISAPPOINTED AND DECIDES TO DISENGAGE FROM THE WENET PLATFORM.

One of the **core functionalities of the WeNet platform and apps is to collect data**. The kind of data collected may differ from one WeNet app to the other. Some data collection may be limited to a specific geographic region, an institution, or a particular interest, while some data collection may expand to WeNet end users worldwide and a variety of life's aspects. The data collected in WeNet will be used to a) improve the WeNet platform by updating algorithms employed in the platform, b) develop new computer models and apps that will run in the WeNet platform, and c) make scientific analyses about the behavior, interests, and opinions of WeNet end users. The **quality of the data matters in all three areas of usage** and will therefore be reflected in this chapter.

Another issue of concern is the **irresponsible use of data published as part of the WeNet Research Infrastructure**. Through the WeNet RI, researchers who are part of the WeNet consortium but also external researchers have access to the data gathered through the WeNet platform. External researchers can gain access upon request. In principle, aligning with democratic goals and the appreciation of open science, anyone should have access to anonymized datasets published as part of the WeNet project. However, this poses the risk that data could be used in low-quality research or be used in the production of false results. In order to prevent such abuses, we discuss the possibility of a researcher code of conduct.

1. ENSURING THE QUALITY OF WENET DATA

When it comes to building computer models, data is key. **Data determines the quality of computer models** that make decisions about the redistribution of resources and knowledge, recommend content to users or personalize user experiences to their individual needs and preferences, and bring together relevant people or items through matching. The quality of the data establishes how well computer models perform, which translates into more or less attractive digital products for people. Furthermore, data determines the quality of scientific analysis, and whether it is appropriate to make general statements about a group or society depends on the representativeness of the data.

Diversity in datasets has become a core topic of contention in the research of dataset construction (Jo & Gebru, 2020). Spectacular cases of incorrect AI decisions or software mistakes are usually referred to in order to **highlight the ethical and social dangers of incorrect forecasts due to inadequate training data**. One prominent example is the study "Gender Shades" by Buolamwini and Gebru (2018), which uncovers the lack of data of female-presenting and darker-skinned faces in training data sets used by big companies to build facial recognition systems. As a result of the skewed datasets, the software developed by popular players in the technology industry misclassified Black women (and colored people in general) with potential consequences of false arrests in law enforcement contexts (Buolamwini & Gebru, 2018; Kantayya, 2020).

The quality of data and datasets is thus important to **prevent discrimination but also to build trust** in AI-based technology. There is the challenge that AI applications necessarily have to work with discrimination in the neutral sense of the word of distinctions/classifications. Appropriate ways must be found here for the revision and criticism of classifications so that they comply with ethical and regulatory norms. Other lines of discussion deal with the relationship between quantity and quality in the generation of training data and emphasize the relevance of normative criteria instead of big data resources (Bender et al., 2021; Jo & Gebru, 2020).

So far, there are no global standards that constitute "good" data practices. It can be expected that the next few years will bring about more aligned **national and international standards for the training, validation, and testing with datasets**. The draft Artificial Intelligence Act (AIA) by the European Commission states in Article 10 that "Training, validation and testing data sets shall be subject to appropriate data governance and management practice" (Artificial Intelligence Act, 2021). Furthermore, Article 10 says that "data sets shall be relevant, representative, free of errors and complete" (Artificial Intelligence Act, 2021). How these standards will be defined remains open at this point. However, they overlap with **core values for collecting, handling, and working with data that are already in use and common practice**.

One reference point for common community practices is the FAIR Guiding Principles by Wilkinson et al. (2016). Together creating the abbreviation FAIR, the principles include "Findable, Accessible, Interoperable, Reusable" data. **Data is findable** if it is labelled correctly, if there is enough information attached (as meta data), and if the dataset or register is searchable. **Data is accessible** if protocols attached to using the data are free and easy to implement; there should be authentication and authorization procedures if necessary. **Data is interoperable** if it is represented in commonly used and understood language and if all relevant references to metadata are attached. **Data is reusable** if it has a clear license for (re)use, is clearly marked as to its origin/source, and adheres to standards in the research community.

Another reference point is the research project KITQAR, which has identified the following principles for ethical data management, see KITQAR Project (2022):

Accuracy Ensure that the data is correct and that individual data points are not included by mistake or compiled accidentally in the wrong area of the dataset.

Completeness Make sure to include all available data in the dataset and avoid large gaps in the dataset. Missing information should be kept to a minimum.

Representational Consistency
Avoid bad formatting. Do not switch between different modes of representation for similar data points.

Timeliness Ensure that all data is relevant and not outdated.

Trustworthiness Do not include external or unfamiliar data in the dataset. If the source of a dataset is unknown, cannot be validated, or seems sketchy, do not use or integrate it with your own data.

2. ETHICAL RESEARCH PRACTICES

This section relates to **preventing scientific misconduct using WeNet data**. Since the WeNet Research Infrastructure is expected to be open to external researchers and data can be received upon request, it can be difficult to anticipate who will use WeNet data and how responsible they will treat it. To prevent scientific misconduct – both intentional and simply by lack of rigor – WeNet should emphasize the importance of ethical research practices and scientific integrity. A **code of conduct for researchers** working with WeNet data should be distributed to external researchers and signed by them upon requesting WeNet data.

Fraud in science occurs more frequently than expected. Usually, researchers seek personal advances from the submission of fraudulent articles. Fraud is spurred by the enormous pressure to collect as many scientific publications as possible known under the term “publish or perish” (Carafoli, 2015, p. 371). Mockery submissions that have grammatical errors in every sentence and were accepted to open access journals are a showcase for the lack of scientific rigor in the scientific and publication industry. The spectrum of scientific misconduct ranges from entirely fabricated papers to results that were ‘forced’ in the sense that the authors had a strong personal bias. Honest mistakes in scientific analysis do not fall under the category of scientific misconduct (Carafoli, 2015).

A common **definition of scientific misconduct** includes the following three violations of ethical practice: **fabrication, falsification, and plagiarism** (Smith, 2006). Fabrication relates to making up research results while falsification constitutes the manipulation of research results. Plagiarism is the theft of parts or entire ideas and/or writing from other people without giving due credit. The definition centering fabrication, falsification, and plagiarism has been developed by the United States Commission on Research Integrity in 1995 (Smith, 2006). There are also broader definitions of scientific misconduct. For instance, the “Norwegian Committee on Scientific

Co-funded by the Horizon 2020 Framework Programme of the European Union

Dishonesty defines research misconduct as ‘all serious deviation from accepted ethical research practice in proposing, performing, and reporting research’" (Smith, 2006, p. 234).

European institutions have released guidelines, principles, and best practices for research ethics and scientific integrity. The **European Federation of Academies of Sciences and Humanities (ALLEA)** is a European network of national academies, among them the German National Academy of Sciences - Leopoldina. Their **Code of Conduct for Research Integrity** was published in 2017 and translated into several languages. On the website of the ALLEA organization, it is stated that "the European Commission recognises the Code as the reference document for research integrity for all EU-funded research projects and as a model for organisations and researchers across Europe" (ALLEA - All European Academies, 2017). Therefore, we will dive deeper into this Code of Conduct and what a WeNet code of conduct for researchers can adopt from this guideline.

The ALLEA Code of Conduct starts by outlining four core principles of research integrity:

- **Reliability** (relates to the quality of research including the soundness of methods, tools, and sources)
- **Honesty** (relates to unbiased positions when conducting research)
- **Respect** (relates to the conduct in the broader research environment, treatment of colleagues, artefacts, research subjects etc.)
- **Accountability** (relates to taking responsibility for a one’s own conduct, methods, and results of the research)

The Code of Conduct (ALLEA - All European Academies, 2017) goes on to outline good research practices. It begins by acknowledging the **importance of scientific context**, i.e. the research environment where a researcher is situated. It is important that research institutions and organizations follow a culture of research integrity, and that training, mentoring, and assistance are provided to the researcher. The Code of Conduct goes on to stress that researchers should always adhere to **state of the art methods** of proper data collection and analysis in their respective fields. The standards of the discipline form a crucial point of reference for the researcher. The Code of Conduct further mentions **data management and the FAIR principles** cited in chapter 5.1. Finally, the Code of Conduct highlights the reciprocity of high-quality science. This means that researchers should draw on each other’s works, cite each other’s works properly, and engage in honest and thorough peer review to support research from other scholars.

In the following, we present a **WeNet draft code of conduct for researchers who work with data gathered through the WeNet platform**. We took inspiration from the ALLEA Code of Conduct (ALLEA - All European Academies, 2017) as well as other sources such as the European Network for Research Ethics and Integrity (ENERI) and the Guidelines for Safeguarding Good Research Practice by the German Research Foundation (2019). The following code of conduct is a first draft that should be discussed with the WeNet consortium. It should further be decided in the consortium whether signing the code of conduct is a prerequisite for researchers to access WeNet data.

Co-funded by the Horizon 2020 Framework Programme of the European Union

WeNet Draft Code of Conduct and Agreement

for researchers using WeNet datasets in their own research
last updated: December 2022

The WeNet platform is an innovation developed by the EU-funded project "WeNet – the Internet of Us." It connects developers, end users, and researchers to leverage the diversity of the community for social engagement, learning, innovation, and research. The WeNet platform is a tool to collect large-scale and high-quality data by adhering to state of the art ethical principles for data protection and legal frameworks such as the European Union General Data Protection Regulation. Researchers using WeNet datasets or researcher-developer teams that collect data via the WeNet platform should follow ethical guidelines for scientific integrity. Ethical research practices with WeNet data are key to upholding trust in the WeNet platform and encouraging end users to engage with the platform while having peace of mind that their privacy rights are respected, and their data contributions serve excellent research.

As a researcher using WeNet data, I hereby declare that I will adhere to the following practices for ethical research:

Following excellence in research. Showing rigor and taking the time to ensure the quality, validity, and authenticity of my research results. Avoiding, whenever possible, to let personal biases influence, manipulate, or change research processes and outcomes.
Denouncing plagiarism. Respecting the intellectual property of peers and citing their works, ideas, and advice properly.
Respecting those who provide data. Always ensuring that data was gathered with the informed consent of data contributors. Following standard anonymization and pseudonymization processes to ensure the data protection rights of data contributors. Preventing re-identification of data contributors. In general, acting in the best interest of data contributors when it comes to their privacy rights.
Handling data with care. Following all WeNet guidelines for collecting, handling, and using data. Ensuring that the data is accurate, complete, relevant, and limited to the research purpose. Ensuring balanced datasets that represent all stakeholders to avoid data bias. Sharing only anonymized data with the public.
Respecting research subjects/interview partners. Showing humanity, respect, cooperation, and interest in the well-being of research subjects. Whenever appropriate, advancing research to the best interest of research subjects.
Accepting complaint mechanisms and taking responsibility for research mistakes. Cooperating with a WeNet ombudsperson for scientific integrity who may receive complaints about the research conducted with WeNet data. Contributing to investigations of alleged scientific misconduct.

Researcher (Printed Name)	Place/Date	Signature

5. MISUSE SCENARIO: DISCRIMINATION THROUGH BIASED DESIGN

Scenario 4: Discrimination of WeNet end users through biased design

WeNet end user A is excited to use an app which is part of the WeNet platform. This app allows users to share short stories as part of a creative writing class at University H. The short stories can be edited and receive feedback from peers. WeNet end user A is disappointed when they note that the app is not optimized for standard screen reader technology. The user is blind and relies on screen reader software to access the short stories shared by their peers and to navigate the app. Frustrated, the end user complains to the owners of the WeNet platform. The platform owners reach out to the app designers and inquire the source of this discrimination. It turns out that the app designers had not considered it necessary to optimize a reading and writing app for blind users. Their ignorance caused suffering for WeNet end user A.

Discrimination through design is a frequented topic in Science and Technology Studies, Critical Design Studies, and Critical Algorithm Studies. Recent years have seen a surge in **studies of technology that reinforces existing inequalities in society** (Eubanks, 2017; Noble, 2018; Wachter-Boettcher, 2017). Such technologies can be AI-based like search, recommendation or decision-making systems that amplify injustices through biased datasets (Zou & Schiebinger, 2018). One example is the 2020 case of alleged fraud by immigrants: "In 2020, it came to light that the Dutch tax authorities had used algorithms that mistakenly labelled around 26,000 parents as having committed fraud in their childcare benefit applications. Many of these parents had an immigration background. They were required to pay back large sums, which led to great financial and psychological difficulties for the families concerned. The data protection authority concluded that the processing of data by the AI system in use was discriminatory" (European Union Agency for Fundamental Rights, 2022).

Technologies that reinforce inequalities can also be apps and interfaces that inlcude or exclude different affordances and thereby include or exclude different groups. One example is accessibility and how screen-reader compatibility makes a huge difference for impaired users (Hamdy, 2020). Another example is sexist and gender bias about bodily form and measurement in the case of the airport scanner (Marzano-Lesnevich, 2019). Here, research has exposed that transgender people are routinely searched in intimate areas of their bodies because their body type does not match an alleged 'male' or 'female' norm (Costanza-Chock, 2020).

Design decisions matter because they immediately affect people. Design determines who can use the product with ease and who may have difficulty in taking advantage of the tool. In some cases, **design decisions can lead to uncomfortable and even offensive experiences for users**. Wachter-Boettcher (2017, p. 6) describes a case where a woman was not granted access to the women's locker room at her gym, because she held an academic title (doctor) and therefore was perceived as male by the system. The design of the locker system with "Dr." being accidentally coded as male does not only cause an unpleasant experience for the user but speaks to a history of gender discrimination in obtaining higher education. The system likely relied on datasets that included only information about men who held an academic title. Preventing such biases in systems is key to ensure equality for users and prevent that historical discrimination is carried on into contemporary practices (see chapter 4.1 on data quality).

While this story relates to a gender bias, another story reveals a **heteronormative bias in the design of a shopping app**. On Valentine's Day, an online shop advertised its products to a female user of the app, saying "Shop Valentine's Day gifts for him" (Wachter-Boettcher, 2017, p. 32). The designers of this special advertisement assumed that female users intend to buy gifts for men and not female partners or friends. This is called a heteronormative bias, where heterosexual relations are expected between two people.

Another example is **racist bias**. The film and photography industry has a history of de-prioritizing darker-skinned customers in the optimization of lighting for (digital) camera pictures. As a result of this, faces of darker-skinned people appeared extra dark due to bad contrast and visibility features (Wachter-Boettcher, 2017, p. 134). In another example, cameras with digital features told Asian customers "not to blink" although their eyes were open. The source of this offense against Asian customers was a facial recognition bias, and the system was not optimized to recognize Asian facial features (Wachter-Boettcher, 2017, p. 135). Further offenses relate to the creation of filters for entertainment purposes. Snapchat was the target of a protest campaign after a filter turned faces in "yellowface," thereby making caricatures of Asian facial features (Levin, 2016). Such racist design features harm minority groups of end users.

Finally, we would like to reiterate the work of former WeNet colleague Hamdy (2020) and their warning of **discrimination against end users with disabilities**. End users who are visually impaired may have difficulty engaging with digital media because much content is text-based or visual. When content cannot be seen by the end user, it is usually read aloud by assistive technologies such as screen readers. "A screen reader is a software application that interprets things on the screen (text, images, links, and so on) and converts these to a format that visually impaired people are able to consume and interact with. Two-thirds of screen reader users choose speech as their screen reader output, and one-third of screen reader users choose braille" (Ashton, 2018). Making apps and services compatible with screen readers is crucial to ensure participation of users with impairment in the digital society. Designers and developers thus need to be aware of the needs of visually impaired users as well as the technical requirements to enable an end user to select their preferred screen reader (Hamdy, 2020).

Discrimination through design can occur if designers hold implicit or explicit bias. Bias can generally take positive and negative form, resulting in favorism in the former case and neglect or denigration in the latter case (Howard & Borenstein, 2018). We are interested in the prevention of **unfair discrimination through negative bias where an already marginalized group of people is further de-prioritized** through design considerations. Explicit bias means that people have a clear priority about whom they would like to advantage and whom they want to exclude. Often, explicit bias is built on racist, sexist or classist ideologies (Benjamin, 2019, p. 61).

Implicit bias is unconscious and unintentional and occurs by internalizing stereotypes through our upbringing and education. An unconscious bias can also result from a person's lived reality: Their experiences in life shape their thinking. People often lack attention to other lived realities because their own experiences take priority. "For example, even though there may be no deliberate intent to cause harm, creating sidewalks without curbs or buildings without elevators can severely disadvantage individuals with movement impairments or disabilities" (Howard & Borenstein, 2018, p. 1523).

Usually, design suffers from a lack of reflection of implicit bias. To prevent unfair discrimination through design, **it is important that designers reflect their own implicit biases and consider other people's lived realities**. What is one's own professional role (in a position of power?), one's position in society (socio-economic standing), one's experiences of privilege and discrimination (Erete et al., 2018). Reflecting implicit bias also includes the reflection of a designer's cultural background as well as the values that stem from this particular cultural framework and which shape the design process (Lazem et al., 2022). Finally, it is important to reflect the potential (unintended) effects of design choices on marginalized communities. Are we accidentally 'othering' groups of end users? This means, are we treating some groups of users as the deserving and rightful users and others as not important? (Hankerson et al., 2016). Answering these questions helps gain a better sensitivity for issues of inclusion and exclusion.

Inspiration for **power-conscious designing that reflects on power relations and norms** is provided by the Design Justice Network. Following from the "Future Design Lab Practice Space" at the ACM 2014 Allied Media Conference, the network has grown to develop ten principles of design justice. The principles are commitments to designing in the best interest of communities and those affected by the design. They see designers as facilitators who work closely with communities to elevate community knowledge, experience, and preferences for design solutions. The principles constitute a shift from the domination of designers' interests to the empowerment of the end users or the local community affected by the design.

In the following, we present a **WeNet draft code of conduct for designers who are creating applications and services through the WeNet platform.** This code of conduct combines design standards from the literature cited above and is geared specifically towards reflecting implicit biases and preventing discrimination through design. The following code of conduct is a first draft that should be discussed with the WeNet consortium. It should further be decided in the consortium whether signing the code of conduct is a prerequisite for designers to use the WeNet infrastructure for developing their own apps.

WeNet Draft Code of Conduct and Agreement
for designers using the WeNet platform to create apps and services
last updated: December 2022

The WeNet platform is an innovation developed by the EU-funded project “WeNet – the Internet of Us.” It connects developers, end users, and researchers to leverage the diversity of the community for social engagement, learning, innovation, and research. The WeNet platform is a tool to develop new applications that serve European and global publics. Ethical design is key to ensure equal participation of end users from all social, cultural, and economic backgrounds in these services. Non-discrimination in alignment with the European human rights framework, including the Charter on Fundamental Rights, is a priority in designing applications and services in the WeNet platform.

As a designer using the WeNet platform, I hereby declare that I will adhere to the following practices for non-discriminatory design:

Denouncing harmful ideologies. Rejecting white supremacy, racism, colorism, colonialism, sexism, homophobia, transphobia, antisemitism, ableism, and other belief systems that build on the subordination of groups of people or lifestyles under others.

Prioritizing the interest of communities affected by the design. Being aware of the community/communities who is/are affected by the design. Surveying their needs and preferences for design solutions. Assessing a priori the (unintended) consequences of implementing the design for the communities. Designing for the benefit and empowerment of the communities affected by the design.

Reflecting one’s own positionality. Being aware of one’s own background and how it influences the design. Making explicit one’s values and interests. Acknowledging positions of power and reflecting on how they impact the design process.

Handling data with care. Following all WeNet guidelines for collecting, handling, and using data that contributes to developing design solutions. Ensuring that the data is representative, and that data bias is ruled out before building design solutions on the data.

Accepting complaint mechanisms and feedback from end users of the design solution. Cooperating with a WeNet ombudsperson for design who may receive complaints about discrimination in WeNet apps and services. Contributing to investigations of alleged discrimination through design features. Showing willingness and openness to improve the design following end users’ feedback.

Designer (Printed Name)   Place/Date   Signature

Co-funded by the Horizon 2020
Framework Programme of the European Union

6. MISUSE SCENARIO: CYBERGROOMING AND CHILD PORNOGRAPHY

Scenario 5: Cybergrooming in WeNet apps

WeNet end user A is using the app "Language training with real natives." This app connects highschool students from different countries to practice a language by communicating in the language in question. WeNet end user A connects with students from England to practice her English. She makes the online acquaintance of WeNet end user B, whose profile says "14 years old, highschool in Birmingham, interested in learning German." They share poems and letters in both languages. At some point, user B asks user A to share a photo of herself. Soon after that, user B asks for naked pictures and user A gets suspicious. Although she feels ashamed, she talks to her parents who contact the WeNet helpline. It turns out that WeNet end user B had faked their account and was a grown man interested in soliciting sexual content from children.

Finally, we would like to include a misuse scenario involving under-age end users of the WeNet platform. The WeNet terms and conditions do not regulate the use of WeNet by children. In other words, children are not prohibited from joining the WeNet platform. They can (legally and technically) create an account and use all functionalities of the WeNet platform.

Particularly concerning is the **potential for abuse of children through communication in the WeNet platform and apps**. A child is understood to mean any person below the age of 18, as per the UN Convention on the Rights of the Child (CRC). In the context of the Covid-19 pandemic, not only leisure activities but also school and extracurricular educational opportunities have shifted to the digital realm. At present, however, there is a general lack of efficient approaches to regulation and empowerment that reliably protect children from dangers online and at the same time support their participation and free development in digital spaces. It is thus necessary to anticipate possible dangers to children.

Since the WeNet platform is a space for learning and social interaction, it might be seen as safe against offenses such as the sharing of child pornography or cybergrooming. However, **security threats often begin at a low level and can then quickly build up and intensify** to the extent that they harm children online but also spill over into their offline world. Especially when the WeNet platform is open to external developers, new apps and services may be created that are geared towards younger audiences. Extracurricula school programs, educational games, and language or maths training may be outsourced to an online tool like a WeNet app. We may also consider younger university students who are starting their higher education at the age of 16 or 17 and use university services via the WeNet platform. These end users fall under the category of children and thus require special protection according to international law and ethical standards.

Co-funded by the Horizon 2020
Framework Programme of the European Union

1. CYBERGROOMING

Cybergrooming is a form of **sexual harassment of children**, whereby an adult (who is not a relative or acquaintance of the child) contacts a child online and seeks to obtain sexualized, pornographic material from the child (Make It Safe). Cybergrooming is a form of online violence against children; it is systematic and sustained, which renders it different from randomized online violence against children. The perpetrator has a clear motive to contact a vulnerable or easily accessible child and uses their experience as an adult to manipulate the child (Wachs, 2014).

While children are nowadays considered 'digital natives,' this does not mean that they are automatically aware of criminal behavior or easily detect manipulation online. In fact, it has been found that children are particularly vulnerable to online manipulation when it comes to online advertisement and product placement (Susser et al., 2019). In terms of social relationships, puberty may contribute to the engagement in risky behavior online. According to Wachs (2014, p. 3), "ICT increasingly serves young people as a medium as well as an experiential ground for their first attempts, testing out, and learning about sexuality" (translated from German original). Children are thus particularly vulnerable to **perpetrators taking advantage of children's drive to explore sexuality in the virtual world**, which is likely perceived as a safe space compared to the 'humiliation' of exposing oneself to conversations about sexuality with parents, friends or teachers.

This is a call to take cybergrooming seriously and protect children from such criminal acts in the WeNet platform. Several steps can be taken:

- The **WeNet platform could introduce an age check** and prohibit the use of WeNet under the age of 18. This is a rather drastic step and would prevent students, schools, and teachers to take advantage of the opportunities that WeNet has to offer. Especially in times of digital learning, the WeNet platform with its strong ethics and data protection should serve as a safe online space that is open to end users of all ages.
- **WeNet apps which are open to children could be closed communities** with a moderator overseeing subscription to or membership in these groups. This way, the moderator has an overview over who is engaging in the community and can halt the participation of unknown or unidentified end users. Given the effort involved on behalf of the moderator, this is a solution for smaller-scale communities.
- The **WeNet platform could offer an educational section** which informs children, teachers, and parents about the dangers of cybergrooming. This section could also include digital literacy training about safe online engagement including how to detect manipulation in computer-mediated and disembodied communication. Educational games could help raise awareness about cybergrooming and teach children to say something if they see that a friend might be affected by criminal initiations.

Co-funded by the Horizon 2020 Framework Programme of the European Union

- The **WeNet platform could designate an ombudsperson for safety**, who can be contacted in case that a WeNet end user suspects criminal activities. There needs to be a protocol for dealing with allegations of child sexual harassment including the notification of authorities.

2. CHILD PORNOGRAPHY

Child pornography is the **visual representation of sexual violence against, between or with participation of children**. There is a distinction between two types of child pornography: voluntary posing of children (e.g. when two teenagers are dating and send each other nude or sexually suggestive photographs) and the forced posing of children (which usually goes hand in hand with child abuse or the worst cases of child pornography such as rape of children) (Huber, 2019, 135ff). Distributing child pornography is illegal, yet case numbers of child pornography on the Internet increase steadily every year. The Internet Watch Foundation (IWF) is a watchdog that searches for and removes child pornography. “In 2021 the IWF took action to remove a record-breaking 252,000 URLs which it confirmed contained images or videos of children being raped and/or suffering sexual abuse” (Internet Watch Foundation, 2021). Most affected are children aged 11-13. The report further states that there was a sharp increase of over 300% of “self-generated” content since the start of the pandemic. “Self-generated” content is such content taped from sexual interactions or posing in front of a webcam, and often involves cybergrooming, see previous section (Internet Watch Foundation, 2021).

Child pornography could be proliferated by malicious actors in the WeNet platform if there are apps that allow for the sharing of images. It can also be possible that child pornography is shared via text with reference to a webpage and url. In what concerns voluntary posing of children, high school students may share nude images of themselves to attract others, which are then copied and proliferated widely. Children may also illegally post nude images of their ex-lovers in an act of revenge after a break-up. Preventing such behavior is very difficult for the WeNet owners and operators. However, several safeguards can be put in place to minimize the spread of illegal content in the WeNet platform. They resemble the measures proposed in response to cybergrooming:

- WeNet could **raise awareness about the danger and illegality of child pornography** and warn of the consequences of posting such content in the WeNet platform. This can be done by posting a statement in a highly visible place in the platform.
- The **WeNet platform could offer links to educational material** which informs children, teachers, and parents about the dangers of becoming victim to child pornography. This section could include material on revenge pornography employed by ex-lovers and how to deal with requests for nude images in teenager relationships.
- The **WeNet platform could designate an ombudsperson for safety**, who can be contacted in case that a WeNet end user suspects criminal activities. There needs to be a protocol for dealing with child pornography including the notification of authorities.

- There are professional organizations that have a history of dealing with child pornography. **WeNet could cooperate with such organizations** to retrieve expertise and knowledge in detecting and removing child pornography in the WeNet platform.

7. SUMMARY RECOMMENDATIONS FOR PREVENTING ABUSE

1. RECOMMENDATION # 1

➢ Develop a concept for content moderation

Communication and social interaction are core features and will only intensify in the WeNet platform. In the different apps that are developed by the WeNet consortium and external designers, WeNet end users communicate with peers from their own communities and broader contexts. Whenever speech is possible, there can be abuse including hate speech.[4] The European human rights framework including the Charter on Fundamental Rights, as well as national legislation in countries of the European Union, demand that abusive communication is identified and removed. Perpetrators should be held accountable, if necessary, by cooperating with law enforcement.

For these reasons, it **is recommended that WeNet develops a sound concept for content moderation in the WeNet platform**. This includes the detection of harmful content and a plan for subsequent handling of such content. Detection of hate speech does not necessarily require automated detection based on a hate speech dataset and natural language processing technology. In fact, these technologies can be biased and reinforce the silencing of marginalized voices (see chapter 2.1). Social mechanisms such as community self-regulation and counter-narratives may be explored. A human in the loop should be available to receive reports of hate speech and act upon such information (see recommendation #4).

2. RECOMMENDATION # 2

➢ Provide transparency about privacy and WeNet's data collection practices

The WeNet platform relies on the engagement of WeNet end users and the collection of their data to provide its services. WeNet end users are thus the essential stakeholders whose well-being and autonomy should be protected and fostered. One important interest of WeNet end users is their right to privacy and data protection. While European and national legal frameworks require data protection measures such as informed consent, these policies are sometimes not clear to the end users because of a lack of information in the platform or app. Transparency is thus key to empowering

[4] While there are legal requirements for the removal of hate speech, there are also *ethical* considerations about speech. Often, legal and ethical ideas of content moderation are intertwined. However, as mentioned in chapter 2, future research should further consider the *ethical* line between acceptable and not acceptable speech under the broad topic of freedom of expression. Current debates about the right of ultra-conservative forces or extremists to express their opinions are inspired by the Twitter takeover of Elon Musk in the United States. But also in Europe, there are extremist groups that turn to the online sphere to 'speak.' What is WeNet's take on where to draw the line? Research on this topic goes hand in hand with an ethically grounded concept for content moderation.

Co-funded by the Horizon 2020
Framework Programme of the European Union

the WeNet end user to make independent decisions about how much and what kind of data they want to provide.

To implement transparency in the WeNet platform, it is recommended to follow two lines of effort: **1) An annual transparency report describing core WeNet practices**. This includes a description of the data collection, the personalization of services with WeNet data, and information about offenses in the WeNet platform. **2) Implementing transparency in WeNet apps**. This means providing "on the go" information about how an app works, e.g., how the collection of data influences its services. This information is integrated into the user flow so that the user does not have to leave the app to find information about data collection or adjust the collection of their data. Inspiration and best practices for both lines of effort are provided in chapters 3.2 and 3.3.

3. RECOMMENDATION # 3

➢ Require researchers and designers in the WeNet community to sign a code of conduct

Researchers in the WeNet community use data from WeNet end users to produce scientific results and establish knowledge. This work must be done ethically and responsibly. Scientific misconduct using WeNet data can lead to a loss of trust among WeNet end users and has implications for the reputation of the WeNet platform. It is **recommended that researchers using WeNet data sign a code of conduct** to declare their agreement to ethical research practices and scientific integrity. A draft code of conduct can be found in chapter 4.2.

Designers in the WeNet community build apps and services by leveraging the infrastructure and tools provided by the WeNet platform. These apps and services should be open to European and global publics and contribute to safe and inclusive learning and social interaction in different communities. To ensure that end users of all backgrounds can participate in the WeNet platform, apps must prevent discrimination and exclusion. Design decisions can lead to discrimination and therefore, it is helpful that designers reflect on their own biases and how they can prevent discrimination in their design solutions. It is **recommended that designers using WeNet infrastructure sign a code of conduct** to declare their agreement to ethical design practices and non-discrimination. A draft code of conduct can be found in chapter 5.

4. RECOMMENDATION # 4

➢ Keep humans in the loop; define designated roles responsible for handling (allegations of) abuse

The WeNet platform – if it continues to expand and invites external researchers and developers to leverage the WeNet infrastructure and pool of end users – it is **recommended that WeNet install ombudspersons responsible for handling complaints from the WeNet community**. These roles should always be filled and be

Co-funded by the Horizon 2020 Framework Programme of the European Union

prominently advertised on the WeNet platform so that stakeholders can easily reach them. The four roles can be filled by the same person if this person is competent in all areas. One open question relates to the ability of WeNet community members to send complaints and concerns anonymously. This option is not available by email and thus the WeNet consortium may want to further discuss the modalities of reaching the ombudsperson.

1. An ombudsperson for content moderation

This person is responsible for dealing with complaints about abusive communication, hate speech, sexualized violence, stalking, and other attacks during social interaction in the WeNet platform. The person is reachable at an email address and, if necessary, cooperates with law enforcement.

2. An ombudsperson for scientific integrity

This person is responsible for managing complaints about research that was conducted within the WeNet platform or by using WeNet data. The person is reachable at an email address. They may discuss allegations of scientific misconduct with the WeNet consortium and bring allegations to the attention of the accused party.

3. An ombudsperson for non-discriminatory design

This person is responsible for managing complaints about unethical design features or discrimination through one of the WeNet apps built on top of the WeNet platform. The person is reachable at an email address. They may discuss allegations of discrimination through design with the WeNet consortium and confront the designers and developers of the app. Further steps may be taken that include sanctions or, in clear cases of outright discrimination, the removal of the app from the WeNet platform.

4. An ombudsperson for children's safety

This person is responsible for dealing with reports of child pornography or suspicious behavior of end users in the WeNet platform. They support the victims of child sexual harrassment in the WeNet platform. The person is reachable at an email address and, if necessary, cooperates with law enforcement.

Co-funded by the Horizon 2020
Framework Programme of the European Union

8. OUTLOOK: ABUSES CONDUCTED BY WENET END USERS

This deliverable has focused on protecting WeNet end users from abuses conducted by researchers, developers, and other end users in the WeNet platform. We highlighted the vulnerable role of end users vis à vis other groups of WeNet community members. We argued that WeNet end users have limited power compared to researchers and developers. This said, WeNet end users' behaviors certainly influence the health of the WeNet community and the quality of data and research coming out of the community. Hate speech and other abusive communication was already discussed in chapter 2. Cybergrooming and inappropriate contact was discussed in chapter 6. Without going into detail, **we wish to point out further abuses conducted by end users that may become more relevant in the future**, when the number of end users grows and more end users outside the pilot sites' student communities engage in the WeNet platform.

1) Fake profiles

One risk of abuse is the creation of fake profiles with good and bad intentions. Some WeNet end users may create fake profiles to protect their identity or to realize a different vision of themselves in the virtual world. While there is no bad intention involved, necessarily, such fake profiles can nevertheless distort datasets and lead to distortions of research results. Creating fake profiles with bad intentions can endanger the safety and well-being of other end users, as in the example of cybergrooming.

2) Anonymous communication

One of the WeNet apps built on top of the WeNet platform, the chatbot 'Ask for Help,' allows for anonymous communication between end users. This feature was installed to encourage exchange on rather sensitive issues, including mental health advice or advice for personal circumstances. However, this feature could also be misused by end users to spread inappropriate content or abusive communication. The WeNet owners and operators then have to balance the protection of the user's expectation of privacy and other end users' rights to physical and mental integrity.

3) Spam and aggressive or manipulative advertisment

Finally, WeNet end users may spread spam via WeNet communication channels such as the chatbot 'Ask for Help.' Advertisement – both obvious and hidden – may influence the WeNet community. This can lead to annoyance among users and user dissatisfaction, and spam can contribute to a bad reputation of the WeNet platform.

Ultimately, the WeNet consortium may want to **consider establishing a code of conduct for users who are engaging in the WeNet platform**. Such a code of conduct would complement the codes of conduct proposed for researchers (see chapter 4.2) and developers (see chapter 5).

Co-funded by the Horizon 2020
Framework Programme of the European Union

REFERENCES

Access Now. (2021). *Transparency Reporting Index*. https://www.accessnow.org/transparency-reporting-index/

ALLEA - All European Academies. (2017). *The European Code of Conduct for Research Integrity*. Berlin. https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction*, *4*(2), 556–579. https://doi.org/10.3390/make4020026

Ashton, C. (2018, December 19). I Used The Web For A Day Using A Screen Reader. *SmashingMagazine*. https://www.smashingmagazine.com/2018/12/voiceover-screen-reader-web-apps/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race after Technology : Abolitionist Tools for the New Jim Code*. Polity Press.

Branley-Bell, D., Whitworth, R., & Coventry, L. (2020). User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases. In M. Kurosu (Ed.), *Lecture Notes in Computer Science. Human-Computer Interaction. Human Values and Quality of Life* (Vol. 12183, pp. 382–399). Springer International Publishing. https://doi.org/10.1007/978-3-030-49065-2_27

Brown, A. (2021, February 26). *What is Hate Speech? Presentation of Hard Cases.* Karlsruhe Institute of Technology. Workshop Hate Speech: What It Is and How It Works, Virtual.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Carafoli, E. (2015). Scientific Misconduct: the Dark Side of Science. *Rendiconti Lincei*, *26*(3), 369–382. https://doi.org/10.1007/s12210-015-0415-4

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2819–2829). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1271

Costanza-Chock, S. (2020). *Design Justice: Community-led Practices to Build the Worlds We Need*. *Information Policy*. The MIT Press.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Daniels, J. (2013). Race and Racism in Internet Studies: A Review and Critique. *New Media & Society*, *15*(5), 695–719. https://doi.org/10.1177/1461444812462849

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. In S. T. Roberts, J. Tetreault, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Third Workshop on Abusive Language Online* (pp. 25–35). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3504

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding Explainability: Towards Social Transparency in AI systems. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. Drucker (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). ACM. https://doi.org/10.1145/3411764.3445188

ENERI. *Website*. https://eneri.eu/

Erete, S., Israni, A., & Dillahunt, T. (2018). An Intersectional Approach to Designing in the Margins. *Interactions*, *25*(3), 66–69. https://doi.org/10.1145/3194349

Eubanks, V. (2017). *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor* (First Edition). St. Martin's Press.

Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Union Agency for Fundamental Rights. (2022). *Bias in Algorithms – Artificial Intelligence and Discrimination*. Luxembourg.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 205395171986054. https://doi.org/10.1177/2053951719860542

German Research Foundation. (2019). *Guidelines for Safeguarding Good Research Practice*. Bonn. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf

Gibert, O. d., Perez, N., García-Pablos, A., & Cuadros, M. (2018, September 12). *Hate Speech Dataset from a White Supremacy Forum*. http://arxiv.org/pdf/1809.04444v1

Gorrell, G., Greenwood, M. A., Robert, I., Maynard, D., & Kalina, B. (2018). Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians. *Association for the Advancement of Artificial*. https://gate-socmedia.group.shef.ac.uk/wp-content/uploads/2019/07/Gorrell-Greenwood.pdf

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The Dark (Patterns) Side of UX Design. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *CHI 2018: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems : April 21-26, 2018, Montréal, QC, Canada / sponsored by ACM SIGCHI* (pp. 1–14). The Association for Computing Machinery. https://doi.org/10.1145/3173574.3174108

Guesmi, M., Chatti, M. A., Vorgerd, L., Joarder, S., Zumor, S., Sun, Y., Ji, F., & Muslim, A. (2021). On-demand Personalized Explanation for Transparent Recommendation. In J. Masthoff, E. Herder, N. Tintarev, & M. Tkalčič (Eds.), *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 246–252). ACM. https://doi.org/10.1145/3450614.3464479

Hamdy, A. (2020). *Diversity and Accessibility in WeNet: Accommodating the Needs of Users with Disabilities.* WeNet - The Internet of Us. https://www.internetofus.eu/wp-content/uploads/sites/38/2020/05/Hamdy-2020-Accessibility-in-WeNet_April2020.pdf

Hankerson, D., Marshall, A. R., Booker, J., El Mimouni, H., Walker, I., & Rode, J. A. (2016). Does Technology Have Race? In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16* (pp. 473–486). ACM Press. https://doi.org/10.1145/2851581.2892578

Heesen, J. (2021). Responsible Freedom: The Democratic Challenge of Regulating Online Media. In L. Trifonova Price, K. Sanders, & W. N. Wyatt (Eds.), *The Routledge Companion to Journalism Ethics.* Routledge.

Heesen, J., Ammicht Quinn, R., Baur, A., Hagendorff, T., & Stapf, I. (2022). Privatheit, Ethik und demokratische Selbstregulierung in einer digitalen Gesellschaft. In A. Roßnagel & M. Friedewald (Eds.), *DuD-Fachbeiträge. Die Zukunft von Privatheit und Selbstbestimmung* (pp. 161–187). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-35263-9_5

Howard, A., & Borenstein, J. (2018). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics*, *24*(5), 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Huber, E. (2019). *Cybercrime*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26150-4

Internet Watch Foundation. (2021). *IWF Annual Report*. https://www.iwf.org.uk/about-us/who-we-are/annual-report-2021/

Jha, A., & Mamidi, R. (2017). When Does a Compliment Become Sexist? Analysis and Classification of Ambivalent Sexism Using Twitter Data. In D. Hovy, S. Volkova, D. Bamman, D. Jurgens, B. O'Connor, O. Tsur, & A. S. Doğruöz (Eds.), *Proceedings of the Second Workshop on NLP and Computational Social Science* (pp. 7–16). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2902

Jo, E. S., & Gebru, T. (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 306–316). ACM. https://doi.org/10.1145/3351095.3372829

Kantayya, S. (Director). (2020). *Coded Bias*. Documentary.

Kasote, S., & Vijayaraghavan, K. (2020). TRUE – Transparency of Recommended User Experiences. In M. Kurosu (Ed.), *Lecture Notes in Computer Science. Human-Computer Interaction. Human Values and Quality of Life* (Vol. 12183, pp. 475–483). Springer International Publishing. https://doi.org/10.1007/978-3-030-49065-2_33

KITQAR Project (2022, January 17). *Kick-Off Meeting: Test and Training Data Quality in the Digital Society.* Verband der Elektrotechnik Elektronik Informationstechnik e. V.; Europa University Viadrina; University of Tübingen; University of Potsdam, Frankfurt am Main.

Kizilcec, R. F. (2016). How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2390–2395). ACM. https://doi.org/10.1145/2858036.2858402

Kleanthous, S., Kuflik, T., Otterbacher, J., Hartman, A., Dugan, C., & Bogina, V. (2019). Intelligent user interfaces for algorithmic transparency in emerging technologies. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (pp. 129–130). ACM. https://doi.org/10.1145/3308557.3313125

Kong, Y., Xie, C., Wang, J., Jones, H., & Ding, H. (2021). AI-Assisted Recruiting Technologies: Tools, Challenges, and Opportunities. In *The 39th ACM International Conference on Design of Communication* (pp. 359–361). ACM. https://doi.org/10.1145/3472714.3473697

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In O. Brdiczka, P. Chau, G. Carenini, S. Pan, & P. O. Kristensson (Eds.), *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126–137). ACM. https://doi.org/10.1145/2678025.2701399

Lazem, S., Giglitto, D., Nkwo, M. S., Mthoko, H., Upani, J., & Peters, A. (2022). Challenges and Paradoxes in Decolonising HCI: A Critical Discussion. *Computer Supported Cooperative Work (CSCW)*, *31*(2), 159–196. https://doi.org/10.1007/s10606-021-09398-0

Levin, S. (2016, August 10). Snapchat Faces Backlash over Filter that Promotes Racist Stereotypes of Asians. *The Guardian*. https://www.theguardian.com/technology/2016/aug/10/snapchat-racist-asian-filter-yellowface

Lumsden, K., & Harmer, E. (2019). *Online Othering: Exploring Digital Violence and Discrimination on the Web*. *Palgrave Studies in Cybercrime and Cybersecurity*. Springer International Publishing. https://doi.org/10.1007/978-3-030-12633-9

Luria, M. (2022). *“This is Transparency to Me:” User Insights into Recommendation Algorithm Reporting*. https://cdt.org/insights/this-is-transparency-to-me-user-insights-into-recommendation-algorithm-reporting/

Make It Safe. *Website on Cybergrooming*. http://www.make-it-safe.net/index.php/de/risiken/risiken-cyber-grooming

Marantz, A. (2019). *Antisocial: Online Extremists, Techno-utopians, and the Hijacking of the American Conversation*. Viking.

Marzano-Lesnevich, A. (2019, April 17). Flying While Trans. *The New York Times*. https://www.nytimes.com/2019/04/17/opinion/tsa-transgender.html

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM*

Transactions on Interactive Intelligent Systems, *11*(3-4), 1–45. https://doi.org/10.1145/3387166

Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, *27*(4), Article zmac010. https://doi.org/10.1093/jcmc/zmac010

Musto, C., Narducci, F., Lops, P., Gemmis, M. de, & Semeraro, G. (2019). Linked open data-based explanations for transparent recommender systems. *International Journal of Human-Computer Studies*, *121*, 93–107. https://doi.org/10.1016/j.ijhcs.2018.03.003

Nassih, R., & Berrado, A. (2020). State of the art of Fairness, Interpretability and Explainability in Machine Learning. In *ACM Digital Library, Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications* (pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3419604.3419776

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Rahman, J. (2012). The N Word: Its History and Use in the African American Community. *Journal of English Linguistics*, *40*(2), 137–171. https://doi.org/10.1177/0075424211414807

Schelenz, L., Segal, A., & Gal, K. (2020). Best Practices for Transparency in Machine Generated Personalization. In T. Kuflik, I. Torre, R. Burke, & C. Gena (Eds.), *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 23–28). ACM. https://doi.org/10.1145/3386392.3397593

Schelenz, L., Segal, A., Gal, K., & Adelio, O. (2022). Transparency in Real World AI-based Systems: User Perceptions and User-centric Design Strategies. *Under Review*.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Shin, D., Lim, J. S., Ahmad, N., & Ibahrine, M. (2022). Understanding user sensemaking in fairness and transparency in algorithms: algorithmic sensemaking in over-the-top platform. *AI & SOCIETY.* Advance online publication. https://doi.org/10.1007/s00146-022-01525-9

Smith, R. (2006). Research Misconduct: the Poisoning of the Well. *Journal of the Royal Society of Medicine*, *99*, 232–237.

Susser, D., Roessler, B., & Nissenbaum, H. F. (2019). Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review*, *4*(1), 1–45. https://doi.org/10.2139/ssrn.3306006

Tintarev, N., & Masthoff, J. (2015). Explaining Recommendations: Design and Evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 353–382). Springer US. https://doi.org/10.1007/978-1-4899-7637-6_10

Vidgen, B., & Derczynski, L. (2020, April 3). *Directions in Abusive Language Training Data: Garbage In, Garbage Out*. http://arxiv.org/pdf/2004.01670v2

Vidgen, B., Margetts, H., & Harris, A. (2019). *How Much Online Abuse is There? A Systematic Review of Evidence for the UK.* The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf

Wachs, S. (2014). Cybergrooming – Erste Bestandsaufnahme einer neuen Form sexueller Onlineviktimisierung. *Enzyklopädie Erziehungswissenschaft Online*.

Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech* (First edition). W.W. Norton & Company.

Wagner, B., Rozgonyi, K., Sekwenz, M.-T., Cobbe, J., & Singh, J. (2020). Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 261–271.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). ACM. https://doi.org/10.1145/3290605.3300831

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). “Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, *15*(2), 87–98. https://doi.org/10.1007/s12193-020-00332-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Yu, L., & Li, Y. (2022). Artificial Intelligence Decision-Making Transparency and Employees' Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort. *Behavioral Sciences (Basel, Switzerland)*, *12*(5). https://doi.org/10.3390/bs12050127

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns (New York, N.Y.)*, *3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

Zhao, R., Benbasat, I., & and Cavusoglu, H. (2019). Do Users Always Want to Know More? Investigating the Relationship between System Transparency and Users' Trust in Advice-Giving Systems. *Proceedings of the 27th European Conference on Information Systems (ECIS))*. https://aisel.aisnet.org/ecis2019_rip/42

Zheng, Y., & Toribio, J. R. (2021). The role of transparency in multi-stakeholder educational recommendations. *User Modeling and User-Adapted Interaction*, *31*(3), 513–540. https://doi.org/10.1007/s11257-021-09291-x

Zou, J., & Schiebinger, L. (2018). Ai Can Be Sexist and Racist - It's Time to Make it Fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8