EUROPEAN UNION AGENCY
FOR FUNDAMENTAL RIGHTS

BIAS IN ALGORITHMS

ARTIFICIAL INTELLIGENCE AND DISCRIMINATION

REPORT

FRA

Bias in Algorithms – Artificial Intelligence and Discrimination

Vienna, 2022

Neither the European Union Agency for Fundamental Rights nor any person acting on behalf of the Agency is responsible for the use that might be made of the following information.

Luxembourg: Publications Office of the European Union, 2022

PRINT	ISBN 978-92-9461-936-5	doi:10.2811/536044	TK0922347ENC
PDF	ISBN 978-92-9461-935-8	doi:10.2811/25847	TK0922347ENN

Photo credits:

Cover: © taa22/Adobe Stock
Page 14: © fizkes/Adobe Stock
Page 16: © Alexander/Adobe Stock
Page 18: © arthead/Adobe Stock
Page 19: © maciek905/Adobe Stock
Page 26: © arthead/Adobe Stock
Page 29: © konstantant/Adobe Stock
Page 30: © pressmaster/Adobe Stock
Page 33: © Kzenon/Adobe Stock
Page 43: © Gorodenkoff/Adobe Stock
Page 47: © ink drop/Adobe Stock
Page 49: © Miha Creative/Adobe Stock
Page 68: © BalanceFormCreative/Adobe Stock
Page 84: © Andrii/Adobe Stock

Foreword

AI is everywhere and affects everyone – from our social media feeds to the social benefits we receive.

Thanks to technology, we receive the information we are most interested in. Administrative decisions can be made more efficiently. Vaccines can be developed faster than before. But as decisions become increasingly automated, it is vital that technology works for us and not against us – or against some of us.

The risks are vast. Think about the now infamous case where hundreds of innocent families were falsely accused of having committed fraud and forced to return social benefits. Many of these families had an immigration background. This was a demonstration of what the use of a biased algorithm can ultimately lead to.

This only underlines that technology needs to be regulated so that we can harness its astonishing potential.

We have to look under the hood and understand how algorithms really work, so we all become more aware of the risks when engaging with it.

In this report, we set out to lift the curtain and show how algorithms work in practice. We developed and tested algorithmic models in the areas of offensive speech detection and predictive policing.

We quickly found that automated hate speech detection is unreliable. Harmless phrases such as 'I am Jewish' or 'I am Muslim' may get flagged as offensive. And yet offensive content may easily slip through.

We also looked at automated predictive policing. And we found that algorithms can generate mistaken information. Essentially sending the police to the wrong parts of the city. This should be a major concern for often under-resourced police forces using this technology.

Our tests highlight how easily algorithms can be biased or develop bias over time. And this can lead to discrimination.

This does not mean that we need to abolish or stop investing in artificial intelligence. It means that humans still have an important role to play. We need to understand how well algorithms work, what the biases are and stay very closely involved in monitoring AI. And we need to always test the applications in the context of their use. In some contexts, algorithms can help us a great deal. In others, they may not at all be fit for purpose.

It is high time to dispel the myth that human rights block us from going forward. More human rights mean more trustworthy technology. More trustworthy technology is a more attractive technology. In the long run it will also be the more successful technology.

If we get this right, I look forward to a future with cures for diseases that are beyond our dreams right now. I look at the delivery of public services to a degree of efficiency and quality which simply is not the case today.

If we move into the right direction, it is astonishing what we could achieve. So, let us work together on getting this right.

Michael O'Flaherty
Director

Contents

Foreword . . . 3

Key findings and FRA opinions . . . 7

CONSIDERING THE DEVELOPMENT OF BIAS IN ALGORITHMS OVER TIME: ‘FEEDBACK LOOPS’ . . . 8

UNCOVERING BIASES IN LANGUAGE PREDICTION TOOLS . . . 11

STRIVING FOR MORE LANGUAGE DIVERSITY . . . 14

INCREASE ACCESS TO RESOURCES NEEDED FOR EVIDENCE-BASED OVERSIGHT OF ALGORITHMS . 15

ENDNOTES . . . 16

1 ARTIFICIAL INTELLIGENCE AND BIAS: WHAT IS THE PROBLEM? . . . 17

1 .1. WHY THIS REPORT? . . . 19

1.2. ARTIFICIAL INTELLIGENCE, BIAS AND FUNDAMENTAL RIGHTS: POLICY CONTEXT . . . 21

1.3. BIAS IN ALGORITHMS AND NON-DISCRIMINATION . . . 22

ENDNOTES . . . 27

2 FEEDBACK LOOPS: HOW ALGORITHMS CAN INFLUENCE ALGORITHMS . . . 29

2.1. EU LAW AND BIAS IN PREDICTIVE POLICING SYSTEMS . . . 31

2.2. PREDICTIVE POLICING IN THE EUROPEAN UNION . . . 34

2.3. SOME CHALLENGES LINKED TO THE USE OF PREDICTIVE POLICING . . . 36

2.4. SIMULATING FEEDBACK LOOPS IN POLICING . . . 38

2.5. ADDRESSING FEEDBACK LOOPS: CONCLUDING POINTS . . . 45

ENDNOTES . . . 48

3 ETHNIC AND GENDER BIAS IN OFFENSIVE SPEECH DETECTION . . . 49

3.1. FUNDAMENTAL RIGHTS AFFECTED BY BIASED SPEECH DETECTION . . . 50

3.2. USING ARTIFICIAL INTELLIGENCE FOR OFFENSIVE SPEECH DETECTION . . . 53

3.3. HOW SPEECH DETECTION ALGORITHMS WERE TESTED FOR BIAS . . . 55

3.4. RESULTS OF ETHNIC AND GENDER BIAS IN OFFENSIVE SPEECH DETECTION . . . 57

3.5. RESULTS OF GENDER BIAS IN OFFENSIVE SPEECH DETECTION . . . 70

3.6. ADDRESSING BIAS IN SPEECH DETECTION: CONCLUDING POINTS . . . 72

ENDNOTES . . . 75

4 LOOKING FORWARD: SHARPENING THE FUNDAMENTAL RIGHTS FOCUS ON ARTIFICIAL INTELLIGENCE TO MITIGATE BIAS AND DISCRIMINATION . . . 77

ENDNOTES . . . 79

5 REFERENCES . . . 81

6 ANNEX I: GLOSSARY . . . 89

7 ANNEX II: ADDITIONAL TECHNICAL DETAILS OF THE SIMULATION OF FEEDBACK LOOPS . . . 91

8 ANNEX III: TECHNICAL DETAILS OF THE OFFENSIVE SPEECH DETECTION ANALYSIS . . . 95

ENDNOTES . . . 101

Figures and tables

Figure 1: Simplistic illustration of a feedback loop . . . 29

Figure 2: Schematic description of the simulated predictive policing process . . . 38

Figure 3: Results of simulation 1: allocation of patrols over time – simple probabilities vs simple machine learning algorithm . . . 40

Figure 4: Source of bias in simplified policing algorithms . . . 42

Figure 5: Predicted crimes per day, with differing historical crime rates per neighbourhood . . . 43

Figure 6: Predicted crimes per day, with historical crime rates reflecting true crime rates . . . 44

Figure 7: Bias mitigation techniques reducing feedback loop formation . . . 45

Figure 8: Average false positive and false negative rates across groups of identity terms, models in English language (%) . . 59

Figure 9: Average false positive and false negative rates across groups of identity terms, models in German language (%) . 60

Figure 10: Average false positive and false negative rates across groups of identity terms, models in Italian language (%) . . . 61

Figure 11: Average predictions of violent speech by selected origins (%) . . . 65

Figure 12: Words' contributions to the model prediction of offensiveness for selected sentences (English, model 3) . . . 67

Figure 13: Words' contributions to the model prediction of offensiveness for selected sentences (Italian, model 1) . . . 67

Figure 14: Likelihood of comments using the African American English dialect and errors in offensiveness predictions . . . 69

Table 1: Predicted probabilities of text being offensive across selected religions/ethnic groups (%) . . . 64

Table 2: False positive rates based on gender used in identity terms in German- and Italian-language models . . . 71

Table A1: Crime data used to inform parameter assumptions . . . 92

Table A2: Characteristics of training data . . . 97

Table A3: Example sentences for the bias test dataset . . . 98

Table A4: Performance metrics of models on test datasets . . . 100

Key findings and FRA opinions

Artificial intelligence (AI)-based algorithms affect people everywhere: from deciding what content people see on their social media feeds to determining who will receive state benefits.

AI technologies are typically based on algorithms that make predictions to support or even fully automate decision-making. Among the main goals of using AI in this way are increasing efficiency and operating systems at a large scale.

But, a central question, and a fundamental rights concern, is what happens if algorithms become biased against certain groups of people, such as women or immigrants?

As research from the EU Agency for Fundamental Rights (FRA) highlights, the use of AI can affect many fundamental rights. While algorithms can be a force for good, they can also violate the right to privacy or lead to discriminatory decision-making, which has a very real impact on people's lives.

A cautionary tale about the possible negative effects of biased algorithms in practice comes from the Netherlands. 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.[1]

Despite the recent sharper focus on the problem of bias in algorithms, this area still lacks a tangible evidence base that employs technical assessments of algorithms in practice and their outcomes. How exactly AI can lead to fundamental rights violations is not yet fully known.

More evidence-based assessments are urgently needed to fill this gap. This report seeks to contribute to this through its focus on 'use cases' of algorithmic modelling. 'Use case' is a term used in software engineering among other fields. This report loosely defines it as the specific application of a technology for a certain goal used by a specified actor. It shows how bias can occur in algorithms and how that bias relates to potential discrimination. It highlights the complexities of bias detection and assessment of potential discrimination, illustrating that bias occurs at several stages, and in different ways.

Although, as this report demonstrates, there is no 'quick fix' for addressing bias in algorithms, it is clear that there needs to be a system for assessing and mitigating bias before and while using algorithms in practice.

Looking at the two use cases in this report, the questions that need to be addressed include the following:

- What if the police are repeatedly sent to certain neighbourhoods based on algorithms, despite faulty crime predictions?
- What if legitimate content posted online by or about certain groups gets deleted more often than others?

This report explores these questions based on original research, combining empirical evidence from the development and testing of algorithms with fundamental rights discussions considering current policy developments. It is based on simulation experiments and analysis conducted to uncover examples of bias in algorithms. Two particular 'use cases' are discussed in the report:

- **predictive policing** – testing a simplified version of a fully automated predictive policing algorithm in relation to crime occurrence in neighbourhoods, focusing on feedback loops (as defined in the following section);
- **offensive speech detection** – the development of algorithms to detect offensive speech, which were then tested for ethnic and gender bias.

The European Commission's proposal of April 2021 for an AI act (AIA) reflects the increased policy and legislative focus on AI.[2] The proposal contains provisions relevant to the protection of fundamental rights. These provisions include requirements for risk management (Article 9), including with respect to fundamental rights, and a conformity assessment for high-risk AI systems (Article 43). Notably, with respect to the focus of this report, the proposed AIA also includes a legal basis for the processing of sensitive data to detect, monitor and correct bias that could lead to discrimination (Article 10 (5)). At the time of writing this report, negotiations around the proposed AIA were ongoing. Fundamental rights protection plays an important role in the negotiations and discussions around the AIA.[3]

The report aims to inform policymakers, human rights practitioners and the general public about risk of bias when using AI, thereby feeding into ongoing policy developments. It is acknowledged that bias can be understood in different ways. This analysis investigates bias in the context of non-discrimination, which is one of the key concerns regarding fundamental rights-compliant AI. The findings pinpoint ways to detect and counteract forms of bias that may lead to discrimination, with the ultimate goal of using AI algorithms in a way that respects fundamental rights.

Test algorithms for bias before and after deploying them, considering their impact over time, and provide guidance on how to collect data on sensitive attributes.

Considering the development of bias in algorithms over time: 'Feedback loops'

The first use case discussed in this report is a simulation of a feedback loop in the area of predictive policing. A feedback loop occurs when predictions made by a system influence the data that are used to update the same system. It means that algorithms influence algorithms, because their recommendations and predictions influence the reality on the ground. This reality then becomes the basis for data collection to update algorithms. Therefore, the output of the system becomes the future input into the very same system.

Any bias in algorithms can therefore potentially be reinforced over time and exacerbated. Feedback loops can lead to extreme results that overestimate realities – so-called runaway feedback loops – which is particularly problematic when applied to 'high-risk' AI applications in the field of law enforcement, for example in the area of predictive policing. The first use case simulates a predictive policing algorithm and demonstrates that several factors can contribute to the formation of feedback loops, including low and varying crime reporting rates, different rates of crime detection and improper use of machine learning. These three factors that contribute to the formation of feedback loops are summarised below.

First, data quality can have an impact on feedback loops. FRA's analysis shows that, for predictive policing, low and varying rates of reporting by victims or witnesses can lead to the creation of biased feedback loops. For example, if predictions of crime rates are based on low reporting rates that fail to reflect the reality of crime occurrence, or the 'true crime rate', this can lead to false predictions and wrong policy decisions.

FRA has repeatedly provided evidence showing low levels of reporting to the police and other authorities in relation to people's experiences of discrimination and crime based on their gender, ethnicity, age and religious background, among other factors. Such reporting rates are influenced by victims' personal characteristics and socio-economic background, as shown in the FRA report on crime, safety and victims' rights[4] and in the FRA report on the results of the second European Union Minorities and Discrimination Survey,[5] the latter of which includes crime reporting rates among different ethnic minority and immigrant groups. This evidence challenges the accuracy of official data sources used for crime predictions.

Second, the detection of different types of crime varies, which can also influence data. For example, some crimes are easier to detect and record. An example of such a crime is car theft, which people have an incentive to report in order to make an insurance claim. Other crimes are not so easy to detect, for example fraud and other financial crimes.

FRA OPINION 1

Users of predictive algorithms need to assess the quality of training data and other sources that influence bias and may lead to discrimination. Such bias and potential discrimination may be developed or amplified over time, when data based on outputs of algorithmic systems become the basis for updated algorithms. Consequently, algorithms that are used to make or support decisions about people, such as predictive policing, need to be assessed before and regularly after deployment. Special attention needs to be paid to the use of machine learning algorithms and automated decision-making.

The EU legislator should make sure that regular assessments by providers and users are mandatory and part of the risk assessment and management requirements for high-risk algorithms.

Certain population groups may be more often associated with crimes that are easier to detect. This may lead to biased predictions over time, as predictions are overly focused on types of crime that are more readily recorded by the police. In addition, the police may behave differently in neighbourhoods that are assumed to have higher crime rates. An increased sense of vigilance among the police in such neighbourhoods may lead to an increase in observed crimes, which can also lead to biased crime records.

Third, in addition to data quality, machine learning algorithms tend to put too much weight on training data. The simulations in the first use case looked at in this report show that a runaway feedback loop (as defined above) occurs more quickly in machine learning models if they are not controlled for overreacting to small signals and differences in the data. As a consequence, the use of techniques for avoiding exaggerated predictions, which mirror training data too strongly (so-called overfitting), are necessary for any algorithm development.

FRA OPINION 2

To better understand how bias can lead to discrimination, data on protected characteristics may need to be collected by users of AI systems to enable assessment of potential discrimination. This data collection needs to be justified, based on strict necessity and should include safeguards in relation to the protection and use of these data. Article 10 (5) of the proposed AIA can provide clarity on the lawful processing of sensitive data that are strictly necessary to detect, monitor and, potentially, mitigate or prevent bias and discrimination. Such a clear legal basis can contribute to better detection, monitoring, prevention and mitigation efforts when using algorithms, but it should be accompanied by appropriate safeguards, including aspects such as anonymisation, pseudonymisation and appropriate limitations with respect to collection, storage, accessibility and retention. Additional implementing guidance on the collection of sensitive data under Article 10 (5) should be considered, notably with respect to the use of proxies and outlining the protected grounds (such as ethnic origin or sexual orientation) that need to be covered.

The development of bias in algorithms over time through such feedback loops risks reinforcing or creating discriminatory practices that affect groups with protected characteristics (such as ethnic origin) disproportionately. To assess potentially disproportionate 'overpolicing' of certain groups, assessments of outputs (algorithmic predictions) are needed with respect to the composition of the target groups.

Ethnic and gender biases in speech detection and prediction models are strong, and need to be assessed when deploying algorithms.

Uncovering biases in language prediction tools

The second use case discussed in this report is a simulation of online hate speech detection systems. Offensive and hate speech detection algorithms, which are currently used in practice, are based on advanced machine learning methodologies and natural language processing (NLP). Given the sheer magnitude of online content, major online platforms have considerably increased their efforts to automatically detect or 'predict' potential online hatred, and have developed tools to do this. However, such tools can produce biased results for several reasons.

Most notably, the level of hatred associated with different identity terms (i.e. words indicating group identities) varies considerably across the data and models that form the basis of the tools. For example, sentences using the term 'Jew' in English-language models lead to a much greater increase in the predicted level of offensiveness than the term 'Christian'. This leads to differences with respect to the predictions of offensive speech for different groups. Those differences can also lead to wrong predictions and classifications.

Several algorithms for offensive speech detection were specifically developed for this report, based on different methodologies and for different languages – English, German and Italian – and subsequently tested for bias. The outcomes show that some terms lead considerably more often to predictions of text as being offensive. For example, in English, the use of terms alluding to 'Muslim', 'gay' or 'Jew' often lead to predictions of generally non-offensive text phrases as being offensive. In the German-language algorithms developed for this report, the terms 'Muslim', 'foreigner' and 'Roma' most often lead to predictions of text as being offensive despite being non-offensive. In the Italian-language algorithms, the terms 'Muslims', 'Africans', 'Jews', 'foreigners', 'Roma' and 'Nigerians' trigger overly strong predictions in relation to offensiveness.

Such bias clearly points to language differences in predictions of 'offensiveness' for different groups by ethnic origin, which means that people who use such phrases are treated differently. Such biased flagging and blocking practices can, for example, lead to differences in access to communication services based on ethnicity. For example, a Jewish person may use the term 'Jew' more often in the online content they post, which may be more readily flagged as offensive and be removed.

One of the reasons for these results is that these terms are strongly linked to online hatred captured in the 'training data' (text datasets including examples of hatred) used for creating the algorithms. The predictions frequently 'overreact' to such terms, flagging text that is usually

FRA OPINION 3

The EU legislator should ensure that assessments of discrimination are mandatory when deploying NLP-based systems such as hate speech detection systems. A context-sensitive and gender-based assessment of potential discrimination is necessary, highlighting potential under- and over-flagging of content. An evidence-driven assessment is needed when testing for bias in algorithms.

The implementation of EU law, such as the Digital Services Act (DSA) and the proposed AIA, should safeguard against discrimination, for example through provisions requiring providers and users of algorithms to provide documentation and carry out assessments in relation to discrimination. With the requirement for increased transparency and assessments of algorithms being the first step towards safeguarding against discrimination, companies and public bodies using speech detection should be required to share the information necessary to assess bias with relevant oversight bodies and – to the extent possible – publicly.

Oversight bodies relevant for protecting fundamental rights, such as equality bodies and data protection authorities, should pay close attention to the potential discrimination in language-based prediction models when exercising their mandates.

not actually offensive as offensive. This is the result of such terms often being considered offensive in training data (i.e. the data used to build the algorithms) – that is, this phenomenon reflects the strong presence of hatred against these groups in the training data.

Furthermore, in the English-language data, the analysis finds a correlation between offensiveness predictions of an algorithm and the likelihood the text uses African American English (which refers to a version of English frequently used by black people in the United States). This correlation may lead to a higher likelihood of African American English text being predicted as being offensive, even if it is not.

Algorithms can also exhibit bias in relation to the gender categories of certain terms. The gender categories of terms were investigated for the German- and Italian-language data, as these languages use gendered nouns. The analysis shows that available language models (pre-trained AI algorithms based on a large amount of text) can lead to gender bias. This bias can lead to differential predictions, for example by considering the feminine version of a term more offensive than its masculine counterpart, or vice versa. For example, the feminine version of 'Muslim' in Italian ('*Musulmana*') is rated by the models more negatively than its masculine counterpart ('*Musulmano*'). This also reflects intersectional hatred, as the rating is based on gender in combination with ethnic origin or religion. This highlights the challenge of using NLP with gendered languages.

These results show how easily bias in relation to protected characteristics can creep into algorithms and can lead to discrimination. Such biases may already exist in off-the-shelf general-purpose AI models, which can be adapted for a specific purpose and are used in fields such as text classification or machine translation. These are based on models that have learned universal language representations based on large amounts of text, and can be adapted for specific purposes, as has been done for this report. Importantly, as the results of the second use case demonstrate, such algorithms cannot be readily used for automated content moderation of hate speech, as they take words linked to protected characteristics – in isolation and out of context – as indications of the presence of offensive speech.

The obligation to respect the principle of non-discrimination is enshrined in Article 2 of the Treaty on European Union (TEU), Article 10 of the Treaty on the Functioning of the European Union (TFEU) (requiring the Union to combat discrimination on a number of grounds) and Articles 20 and 21 of the EU Charter of Fundamental Rights (equality before the law and non-discrimination based on a non-exhaustive list of grounds). All prohibited grounds of discrimination as listed in the Charter are relevant when it comes to the use of algorithms.

More specific and detailed provisions in several EU directives also enshrine this principle, with varying scopes of application. The proposed equal treatment directive[6] would provide even more protection for several grounds of discrimination, as it fills specific gaps.

FRA OPINION 4

The EU's anti-discrimination legislation is crucial for safeguarding a high level of equality in the EU. The present analysis shows that speech algorithms include strong bias against people based on many different characteristics, such as ethnic origin, gender, religion and sexual orientation. As a consequence, the EU legislator and Member States should strive to ensure consistent and high levels of protection against discrimination on all grounds, including (at a minimum) sex, racial or ethnic origin, religion or belief, disability, age, sexual orientation, gender identity and gender expression in different areas of life. This discrimination should be tackled using various existing laws that safeguard fundamental rights. In addition to non-discrimination legislation, existing data protection laws should also be used to address non-discrimination regarding the use of algorithms for decision-making.

The requirements for high-risk AI use cases – as included in the proposed AIA – should increase transparency and allow for the assessment of discrimination of algorithms. This information, with respect to AI use cases, can be used to enforce existing non-discrimination and data protection laws.

Finally, equality bodies should step up their efforts to address discrimination complaints and cases linked to the use of algorithms. In order to do this effectively, they should employ specialised staff and cooperate with data protection authorities and other relevant oversight bodies.

There is a need to promote language diversity in tools available for natural language processing.

FRA OPINION 5

The EU and its Member States should consider measures to foster more language diversity in NLP tools as a way of mitigating bias in algorithms and improving the accuracy of data. As a first step, this should include promoting and funding NLP research on a range of EU languages other than English in order to promote the use of properly tested, documented and maintained language tools for all official EU languages.

The EU and its Member States should also consider building a repository of data for bias testing in NLP. Such a repository should conform to EU standards of data protection, contain high-quality data in all EU languages to enable testing for biases and be continually updated and maintained.

Striving for more language diversity

While NLP technologies and the availability of related tools for English have improved considerably in recent years, NLP tools for other languages are lagging far behind. This report uncovered a clear imbalance between the tools and knowledge available for NLP technologies for English and those available for other languages. For EU languages, such as German and Italian, results revealed the quantity and quality of available data and NLP tools to be insufficient, resulting in reduced performance of the algorithms detecting offensive speech, while also exhibiting strong levels of bias.

There is a need to increase knowledge, awareness and resources for bias testing of algorithms.

Increase access to resources needed for evidence-based oversight of algorithms

The differences in bias in algorithms found in this project differ not only across languages but also across the tools and methodologies used. Some AI models are based on algorithms that were developed for general language detection and prediction tasks based on large bodies of text. Such 'pre-developed' models are needed for speech detection algorithms in many cases. The research found that bias differed depending on which tools were used. This means that bias is already embedded in pre-developed general-purpose speech models, which are often developed by large companies with access to vast amounts of data and computing power. Assessing and documenting bias in such pre-developed models is challenging in the absence of full documentation and available tests for identifying bias in such tools. In addition, datasets are often difficult to obtain. This is partly because NLP researchers are overly cautious and avoid sharing data, often because they lack knowledge of data protection rules.

FRA OPINION 6

To increase the application of trustworthy AI, compliant with fundamental rights, more EU and national funding for fundamental rights assessments of existing software and algorithms is needed to support studies of available general-purpose algorithms. This would help deployers and users of AI tools to more easily conduct their own fundamental rights impact assessments before and during the use of certain AI systems.

The EU and its Member States should improve access to data and data infrastructures for identifying and combating the risk of bias in algorithmic systems. This includes ensuring access to data infrastructures for EU-based researchers. This could be achieved through investment in cloud computing and storage infrastructures, designed in accordance with EU standards for data protection, software safety and energy efficiency. EU-based researchers should be granted access to such infrastructure to foster public scrutiny.

In this respect, Article 40 DSA allows for researchers to better access data from online platforms. This article should be used to the extent possible, without bureaucratic obstacles, to allow easy and widespread access to data needed for – the sole purpose of – bias- and discrimination-related research on online platforms' conduct.

To further improve availability of evidence of bias, the European Commission, the European Data Protection Board and the European Data Protection Supervisor should look at the need to address issues of correctly implementing data protection law in relation to data sharing with respect to sensitive data for the purpose of researching and monitoring discrimination. Without clearer guidance, misinterpretation of data protection law may unnecessarily stand in the way of independent evidence-based oversight of the risk of bias in algorithms.

Endnotes

[1] Dutch Parliamentary Committee (2020), p. 14.

[2] **Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence** (Artificial Intelligence Act) and amending certain Union legislative acts, COM(2021) 206 final, Brussels, 21 April 2021.

[3] European Parliament (2022), amendment 89.

[4] FRA (2021a).

[5] FRA (2017).

[6] Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation (proposed equal treatment directive), COM(2008) 426 final.

1

ARTIFICIAL INTELLIGENCE AND BIAS: WHAT IS THE PROBLEM?

Algorithms are used to automate many tasks that affect our lives, often on a scale that cannot be matched by work undertaken by humans. Algorithms that can predict, label, analyse and recommend have opened up new horizons and can support decision-making across many domains. They hold the promise of huge benefits to the economy and society at large.

Alongside benefits, however, algorithms also pose risks to fundamental rights. Therefore, it is essential for those responsible for the development and use of algorithms to examine their possible impact on people and their fundamental rights. This balancing act between innovation and fundamental rights is also at the heart of EU policy efforts to regulate AI and related technologies.

FRA has pointed out in previous reports on AI and big data that only a rights-based approach guarantees a high level of protection against possible wrongdoing related to new technologies.[7] One such potential wrongdoing – and the subject of this report – involves the risk of bias in algorithms: the tendency for algorithms to produce outputs that lead to a disadvantage for certain groups, such as women, ethnic minorities or people with a disability.

This is not an imaginary problem, but one with real-world effects. In the Netherlands, for example, the tax authorities used algorithms that mistakenly labelled around 26,000 parents as having committed fraud in their childcare benefit applications. A disproportionate number 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.[8] The scandal has become a cautionary tale about the impact on people of decisions made (or supported) by biased algorithms.

How could biased algorithms affect your day-to-day life?

Imagine living in a neighbourhood with a very strong police presence and being stopped and searched on the streets, perhaps frequently. Are the reasons behind the police presence based on correct and representative facts, and what role – if any – does the police's new AI-based crime prevention tool play in this? Does bias play a role?

Imagine you are from a certain religious community and an active social media user, but several posts mentioning your religion are taken down by the platform as potentially offensive. Is this something people with a different (or no) religious background also experience, and to what extent?

So how do these AI systems/algorithms become 'biased'? The general answer is simple enough: the bias itself and the resulting discrimination is pervasive in society, rooted in psychological, social and cultural dynamics, and hence reflected in the data and texts that are used for developing AI models. The use of algorithms to support decision-making processes is often portrayed as rational and neutral. But machines and technology are not neutral, because they are developed and used by humans. Where bias is present in human decision-making, it may be transferred to machines.

The question of exactly how bias creeps into the predictions of algorithms – the topic of this report – requires a more complex answer. The two situations mentioned in the box above – being stopped by the police and having social media posts about one's own religion deleted – are central to the case studies in this report. Algorithms used for predictive policing and for offensive speech detection were tested to clearly illustrate how and where bias may occur and when it leads to discrimination.

Algorithms, models and artificial intelligence

The term 'algorithm' is widely used in the context of big data, machine learning and AI. In computer science, an algorithm is a sequence of commands for a computer to transform an input into an output. A sequence of commands to sort a random list of people according to age is one example of an algorithm. In that example, we provide a computer with the random list (input), execute the algorithm (commands) and the computer produces a list sorted by age (output).

Algorithms are often used to make predictions, for example predictions regarding the profile of people who are likely to buy a certain product, the weather forecast or spam detection. For a specific task, an algorithm is fed with data; this creates a **model** that is used in practice for a real-world task. Following common machine learning terminology, an algorithm is the 'raw' state, and the model is what you get when the algorithm has been 'trained' on data.

The term '**artificial intelligence**' is more difficult to define. 'AI' does not refer to a single, tangible thing but to current technological developments and processes in general. Most of what is discussed under the umbrella of 'AI' refers to the increased automation of tasks through the use of machine learning and automated decision-making. At the core of current AI discussions and machine learning applications lies the use of algorithms. FRA's *Getting the future right* report* includes a broader discussion of AI and related terms. Efforts to clarify the definition of AI include work by the European Commission's High Level Expert Group on Artificial Intelligence and the OECD.**

* *FRA (2020),* **Getting the future right – Artificial intelligence and fundamental rights***, Luxembourg, Publications Office.*

** *OECD (2022),* **OECD Framework for the Classification of AI systems***, OECD Digital Economy Papers, No. 323, Paris, OECD Publishing.*

1.1. WHY THIS REPORT?

At the time of writing, the EU is striving to become the first regional actor to regulate AI using a dedicated legal instrument. The EU has set out to play a "leading role in setting the global gold standard"[9] when it comes to addressing the risks generated by specific uses of AI, while at the same time assisting EU Member States in their transitions to digital societies, including promoting the development and uptake of AI in the EU.

As FRA stressed in its *Getting the future right* report,[10] only concrete examples allow for a thorough examination of whether, and to what extent, using technology interferes with fundamental rights and whether any such interference can be justified, in line with the principles of necessity and proportionality. Such concrete examples need to be evidence based. While availability of evidence and data on fundamental rights implications of using algorithms and AI has increased in recent years, there is still a lack of evidence in relation to how and where bias occurs and when it leads to discrimination.

A recent study by the European Parliamentary Research Service emphasised that, as AI is created by humans, it can be susceptible to bias. This bias "most frequently occurs when machine learning applications are trained on data that only reflect certain demographic groups, or which reflect societal biases."[11] The same study highlights the additional challenge that AI applications are often 'black boxes', making it impossible for the consumer to judge whether the data used to train them are fair and representative, which in turn makes biases hard to detect and mitigate.[12]

Determining where bias comes from is therefore a challenging exercise. In addition, there are many different kinds and notions of biases, not all of which are necessarily negative or harmful (see Section 1.2). As FRA's previous reports have highlighted, algorithms are only as good as the data that are used to develop them.[13] In practice, poor-quality and/or biased data often all too easily slip into algorithmic processes, rendering them flawed and potentially discriminatory. There are also other sources of bias. Algorithms may be trained on data that are not necessarily biased, but are unrepresentative. In this case, the data used for such algorithms cannot be used to make generalisations about other groups. For example, if a face detection algorithm is trained using predominantly male faces, its predictions might be problematic when applied to female faces.[14]

This report looks under the hood of algorithms in the context of two particular 'use cases' (see the next section) to demonstrate some sources of bias. The report also firmly places the discussion of bias in the context of non-discrimination. Discrimination is a negative result of bias, which needs to be prevented. Regarding this, **FRA adds unique findings from original applied research to underpin the much-needed body of evidence on bias and AI**, while suggesting concrete steps to counter such bias. As questions of how bias can be detected and counteracted in AI applications are at the centre of discussions around regulating AI, these findings can help point policymakers to potential safeguards and mitigation strategies to avoid unwanted bias and discrimination in AI technologies.

This report is intended to be read by policymakers working on AI and providers and users of AI. The results of the report are also relevant to human rights practitioners and academics dealing with the topic of new technologies and fundamental rights. Finally, the report informs the general public about the risks of bias when using AI.

What has FRA done in this area to date?

FRA's first report on the subject of AI was published in 2018.* It highlighted the fact that algorithms are only as good as the data they are fed. If the data are outdated, incorrect, incomplete or poorly selected, results too will be questionable. With endless amounts of data being so quickly produced on the internet, the lack of quality control regarding how these data are produced and then used is a serious concern.

A further FRA focus paper published in 2019** reemphasised the risk that AI systems based on incomplete or biased data can lead to inaccurate outcomes that infringe on people's fundamental rights, including non-discrimination. It also explored the importance of using high-quality data to ensure high-quality algorithms, including the need for developers and providers of algorithms to be transparent about which data are used in AI systems in order to help prevent possible fundamental rights violations.

As one highly relevant area of AI, fundamental rights considerations in the context of law enforcement in relation to facial recognition technology were dealt with in a 2019 FRA report.*** The report provides an overview of what the technology is, how law enforcement authorities (plan to) use the technology and what the main fundamental rights concerns are in relation to its use.

In a 2020 report, FRA highlighted some concrete examples of bias in algorithms.**** In addition to highlighting the need for further studies of potential discrimination resulting from the use of AI systems, some professionals FRA interviewed for this report underscored that results from complex machine learning algorithms are often very difficult to understand and explain. This leads to the conclusion that further research to better understand and explain such results ('explainable AI') could also help to better detect discrimination when using AI.

* *FRA (2018a),* **#BigData: Discrimination in data-supported decision making**, *Luxembourg, Publications Office.*

** *FRA (2019),* **Data quality and artificial intelligence – Mitigating bias and error to protect fundamental rights**, *Luxembourg, Publications Office.*

*** *FRA (2019),* **Facial recognition technology: Fundamental rights considerations in the context of law enforcement**, *Luxembourg, Publications Office.*

**** *FRA (2020),* **Getting the future right – Artificial intelligence and fundamental rights**, *Luxembourg, Publications Office. See Section 4.5 for more information on some of the main challenges associated with discrimination and the use of AI.*

What does this report cover?

FRA gathered evidence for this report by engaging with two practical examples of the use of algorithms in order to get a better understanding of the extent of potential bias, and to assess the legal implications, particularly with respect to fundamental rights. The overall design of the research was developed by FRA and conducted by a consortium of data scientists and lawyers led by Rania Wazir. Based on the findings, FRA carried out further analysis and drafted this report. The two use cases covered the following examples.

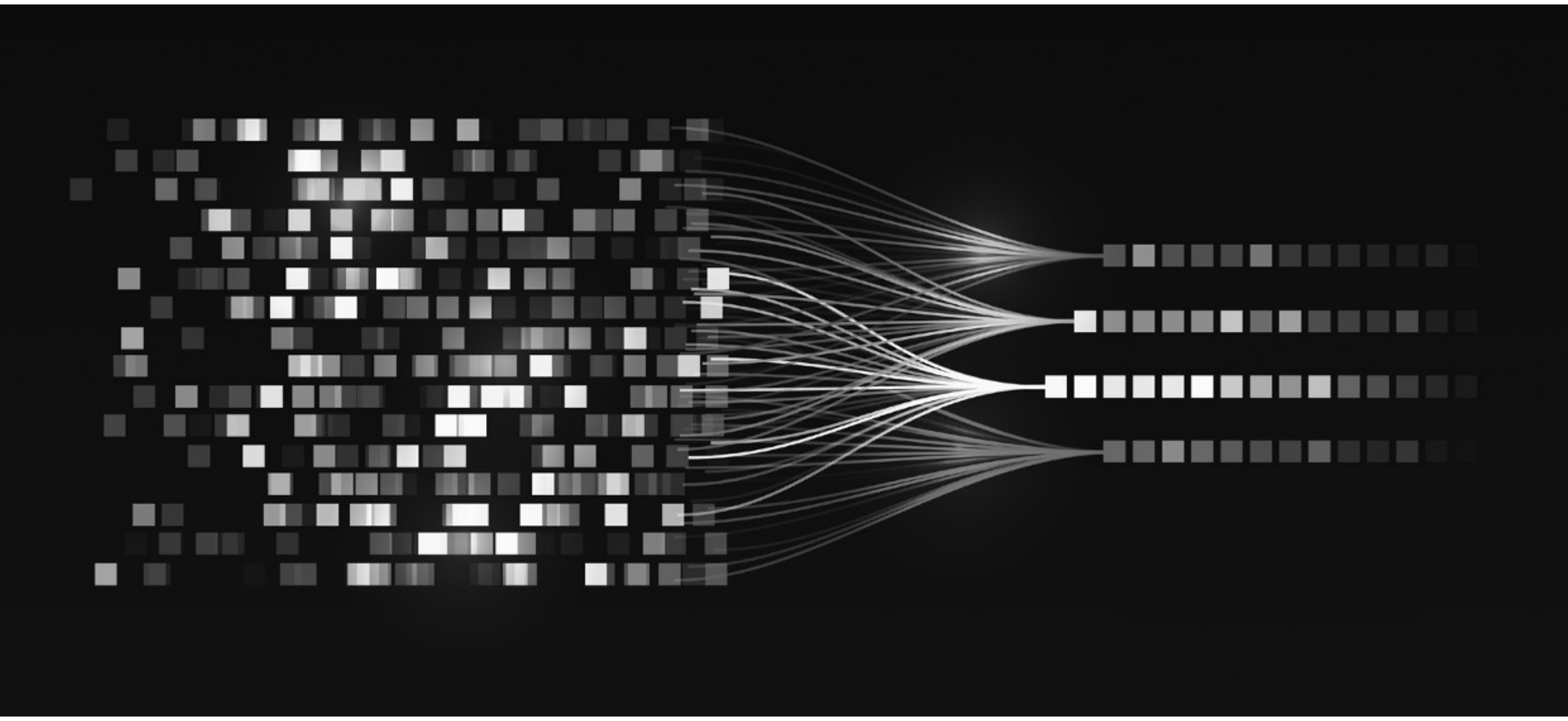
The first example (elaborated on in Chapter 2) analyses how algorithms can exacerbate bias over time, often referred to as the formation of runaway **feedback loops**. It is based on a simulation study applied to the example of **predictive policing** contexts. A feedback loop occurs when decisions based on predictions made by an AI system influence the data that are then used to retrain or update the system. So-called runaway feedback loops not only perpetuate biases in the data, but can also actually increase them. For example, if police forces are advised to monitor one area based on predictions that are influenced by biased crime records, then police will detect more crime in that area.

The second example (elaborated on in Chapter 3) analyses ethnic and gender **biases in offensive speech detection algorithms**. Algorithms were developed based on real-life offensive speech datasets, using different approaches, including pre-trained AI models. The models were developed for English-, German- and Italian-language datasets. These models were then tested against invented phrases to see how the predictions change for different terms related to ethnic groups and gender. For example, the sentence 'I hate […]' was used, where '[…]' was populated with various terms for groups (e.g. African, European) to see which terms trigger more offensive predictions.

Chapters 2 and 3 present the main findings that emerged from the analysis of these two use cases. The remainder of this chapter presents a brief outline of the policy context and legal framework relating to biased algorithms and non-discrimination.

1.2. ARTIFICIAL INTELLIGENCE, BIAS AND FUNDAMENTAL RIGHTS: POLICY CONTEXT

EU institutions have become increasingly engaged in the area of AI, bias and other fundamental rights considerations with respect to policy positions and, more recently, legislative proposals. In 2017, a European Parliament resolution on fundamental rights implications of big data called for the strong enforcement of fundamental rights in relation to new technologies.[15] The European Council called for a European approach to AI and emphasised in its strategic guidelines for 2019–2024 the need to "ensure that Europe is digitally sovereign" and for policy to be "shaped in a way that embodies our societal values".[16] The European Commission published its 2018 communication on AI for Europe,[17] established the High Level Expert Group on Artificial Intelligence and committed to putting forward legislation "for a coordinated European approach on the human and ethical implications of [AI]".[18] In February 2020, the European Commission published a white paper on AI, setting out policy options for meeting the twin objectives of "promoting the uptake of AI and addressing the risks associated with certain uses of this new technology".[19]



On 21 April 2021, the Commission published its proposal for an AIA.[20] Negotiations on the AIA proposal were ongoing in the European Parliament and Council at the time of writing this report. The AIA proposal – with provisions regulating different risk categories of AI applications – forms part of the European Commission's **digital strategy**, and is a key step in the EU's endeavour to make its law fit for the digital age. The AIA proposal is part of a tranche of proposals that must be understood in tandem, including the DSA (with provisions on recommender systems and a risk assessment in relation to fundamental rights, expressly mentioning the prohibition of discrimination, by very large online platforms), the digital markets act (with provisions on AI-relevant hardware, operating systems and software distribution), the draft machinery regulation (revising the Machinery Directive[21] in relation to AI, health and safety, and machinery), the draft data governance act (concerning data-sharing frameworks) and the revision of the Product Liability Directive[22] in relation to AI.

1.3. BIAS IN ALGORITHMS AND NON-DISCRIMINATION

Biases in algorithmic systems may lead to discrimination. They also have the potential to amplify discrimination because of their potential scale of application or because of feedback loops.[23] However, it is important to distinguish between bias and discrimination.

Policy processes at the international level

Globalpolicy.AI is a platform involving eight intergovernmental organisations with complementary mandates that cooperate to help policymakers and the public navigate the international AI governance landscape and access the necessary know-how and tools to inform AI policy development.

The Globalpolicy.AI initiative launched in September 2021. The intergovernmental organisations currently involved are the Council of Europe, the European Commission, FRA, the Inter-American Development Bank, the Organisation for Economic Co-operation and Development (OECD), the United Nations, the United Nations Educational, Scientific and Cultural Organization and the World Bank Group.

It highlights international efforts such as the **Council of Europe's efforts to create an international instrument to regulate AI** and **UNESCO's recommendation on the ethics of AI**.

For more information, see the ***Globalpolicy.AI website****.*

Definitions of bias: Not to get lost in translation

The term ‘bias’ can have a different meaning depending on the context in which it is used and the particular discipline it comes from, for example law or computer science. It is therefore important to clarify its meaning in the context of this report. Bias can refer to any of the following.

- **Differential treatment based on protected characteristics, such as discrimination and bias-motivated crimes.** This refers to an inclination for or against a person or group based on protected characteristics, such as ethnic origin, gender, religion, colour or sexual orientation. Discrimination defines a situation in which an individual is disadvantaged in some way on the basis of ‘one or multiple protected grounds’. Crimes committed with a bias motivation are a particularly severe example of a result of biases against people based on their (assumed) characteristics.* Such definitions are often used in legal contexts and the social sciences.
- **Differentiation.** Bias understood in this sense is necessary for the proper functioning of a statistical or machine learning algorithm. For example, a machine learning model that has to differentiate between oranges and pears has to have bias towards labelling round, orange objects as oranges. Such use of bias is mainly found in computer science and machine learning.
- **Statistical bias.** This refers to the systematic difference between an estimated parameter and its true value. Statistical bias exists when data are not adequately measuring what they are intended to measure. For example, gross domestic product per capita is not necessarily a good measure of the standard of living in a country, as it does not account for inequality of income distribution. In addition, data and the resulting statistical estimates may not be representative of the target population. For example, if a sample of the general population contains more men than women, it is said to be biased towards men. Bias is mainly understood in this way in statistics.
- **Offset from origin.** In the context of deep learning, bias is also the name for an estimated parameter. The fixed number indicating the average baseline estimate in the linear weight functions of neural networks is called bias; it is often referred to as a ‘constant term’ or ‘intercept’ in classical regression analysis. It is a purely technical term, and as such it is not relevant to the present discussions, although it is used in neural networks.

Bias is analysed in the context of discrimination (as a legal and normative concept) in this report. Discrimination is mainly linked to prejudices picked up or enshrined in data, but may also be the result of statistical bias.

* *FRA (2018),* **Handbook on European non-discrimination law**, *Luxembourg, Publications Office, Chapter 2.*

```
this.state = {
    page_type: !1
}, this.views = [], this.overlays = [];
var t = this;
this.pages = {}, this.app_settings = {}, this.dynamic_filters_ob
    return {
        name: "",
        rendered: !1,
        initialized: !1,
        filters_options: [],
        filters: []
    }
}, this.init = function(a) {
    var r = this;
    r.app_settings = $.extend({
        api_url: "",
        __ROOT__: "",
        trigger: "#13251231",
```

First, not all forms of bias relate to protected characteristics. For example, an algorithmic model that differentiates between people based on whether or not they have a pet does not directly target a protected characteristic, and having a pet or not is unlikely to act as a proxy for a protected characteristic. Second, even if an algorithm contains bias related to a protected characteristic, the result may still not be discriminatory if the decision taken based on the algorithmic system does not lead to less favourable treatment, or is justifiable for the purpose it is employed. For example, an algorithm that chooses which song from a predetermined playlist you hear next may have a gender bias, but it is questionable whether presenting songs in a different order constitutes less favourable treatment. Algorithms may also contain bias related to protected characteristics, which are justifiable in relation to genuine occupational requirements, such as age limits for certain jobs requiring physical fitness.[24] Therefore, to determine the legal implications, an algorithm must always be assessed within the particular context and purpose of its use.

Forms of discrimination

Discrimination can take different forms:

- **direct discrimination** takes place when a person receives less favourable treatment than another in a comparable situation, based on a protected ground;
- **indirect discrimination** takes place when an apparently neutral provision, criterion or practice puts people with a particular protected characteristic at a disadvantage compared with others;
- when several grounds of discrimination are involved, **multiple discrimination** (where the grounds operate separately) or **intersectional discrimination** (where the grounds interact and are inseparable) can occur;
- **discrimination by association** is where a person is treated less favourably based on another person's protected characteristic, but is not themselves the person with the protected characteristic.

For more information, see FRA (2018), **Handbook on European non-discrimination law***, Luxembourg, Publications Office.*

Bias in algorithms may lead to direct discrimination when reliance on a protected characteristic leads to less favourable treatment. This will normally occur only where coded parameters and/or training data and input data include features that directly indicate a protected characteristic (e.g. where a predictive policing algorithm includes information on the ethnicity of residents in a particular neighbourhood or where a content moderation algorithm contains information about the ethnic origin of the author of a particular post). Such information can be easily spotted when it is directly included, and this allows for the assessment of differential treatment based on those characteristics.

More often than not, however, algorithmic bias leads to indirect discrimination because of the inclusion of proxies. A proxy is a seemingly neutral piece of information that is nevertheless strongly related to a protected characteristic. For example, shoe size as a proxy for gender or names as a proxy for ethnicity. Discrimination resulting from the use of proxies is more difficult to prevent, as there is a potentially limitless number of proxies, and their correlation to a protected characteristic will be evident to various extents. For example, the selection of certain neighbourhoods for enhanced policing activities may correspond to neighbourhoods that are composed mainly of certain minorities. While geographical area may not be a reason for discrimination, the composition of the population of the neighbourhoods may be.

Where someone brings a claim of discrimination (e.g. a user of a social network whose posts tend to be removed claims discrimination based on ethnicity), the burden of proof usually rests upon that person to establish a presumption of discrimination.[25] Statistical evidence, for example based on discrimination testing, can be useful for this purpose. There is no generally accepted rule as to what kind of statistical bias triggers the presumption of discrimination. While some scholars have proposed certain formalised schemes to assess the disadvantages across demographic groups (called conditional demographic disparity) as a starting point to assess potential discrimination, it is also well established that the most appropriate fairness metrics arguably always depend on the specific context.[26] While it is clear that a rule that puts 85 % or 90 % of users with a particular ethnicity at a disadvantage will trigger a presumption of discrimination,[27] the threshold of what counts as sufficiently significant discrimination cannot be defined in the abstract but must be assessed on a case-by-case basis, taking into account a number of factors.[28]

When a presumption of discrimination has been established, the burden shifts to the alleged 'defendant', who must provide evidence that the less favourable treatment was not based on a protected characteristic.[29] A presumption of discrimination can be rebutted by the defendant. To do so, they must prove either that the victim is not in a similar situation to their 'comparator' or that the difference in treatment is based on an objective factor unconnected to the protected ground. This could be done by providing insight into the real source code, by applying other (and at least as appropriate) fairness metrics[30] or by demonstrating how the algorithm works with the help of post hoc explanation tools showing that the dependency on the protected characteristic does not exist.[31]

Finally, some scholars suggest that algorithmic systems may discriminate not only along the lines of existing human prejudice and discriminatory behaviour, but also based on new grounds such as profiling identities based on a combination of behavioural and demographic characteristics.[32] As far as these categorisations serve as proxies for existing protected characteristics, such as gender or race, any 'new combinations' are already covered by existing European non-discrimination law. If, however, algorithmic systems treat individuals less favourably because of (a collection of) characteristics that bear no link to the existing protected characteristics, new groups of disadvantaged individuals may emerge. As a result, some scholars argue that the prevention of discrimination linked to legally protected grounds may no longer be sufficient. Such groups could, for example, comprise people who play online games, dog owners or 'sad teenagers'.[33]

The fact that AI and algorithmic systems often lack transparency complicates the detection of discrimination, as previously noted by FRA.[34] For example, proving a direct form of discrimination in a fully disclosed relatively simple machine learning algorithm or statistical model (such as logistic regression) may be relatively straightforward. Such algorithms may directly indicate the extent to which predictions changed based on protected characteristics (such as gender). However, detecting and proving discrimination becomes more difficult if the algorithm is not fully disclosed or is more complex. Still, evidence for discrimination could then be sought by running tests on the systems using experimental methods. If the output of algorithms differs when only information on protected characteristics changes, there is a risk for discrimination. As a consequence, a variety of methods need to be developed and deployed to investigate potential bias, to address potential discrimination.

Moreover, in the absence of opportunities to test algorithms using experimental methods, other information can help indicate discrimination, such as the use of so-called explainable AI technologies, showing which information in the data contributes most to the predictions.

Data protection and measuring discrimination in artificial intelligence

In its 2018 report ***#BigData: Discrimination in data-supported decision making***, FRA highlighted that the processing of personal data related to protected characteristics may be needed to detect and potentially mitigate discriminatory outcomes when using algorithms. Such data collection, however, needs to comply with data protection law, which includes specific requirements for data linked to protected characteristics.

In general, the processing of special categories of personal data revealing, among other things, racial or ethnic origin, religious beliefs or sexual orientation is prohibited except for the grounds of justification listed in Article 9 (2) General Data Protection Regulation (GDPR). The two grounds with the broadest scope of application are detailed in points (a) (the data subject's explicit consent) and (g) (necessity for reasons of substantial public interest, on the basis of EU or national law). The latter must be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

Detecting algorithmic bias or discrimination is not a specifically mentioned justification for processing sensitive personal data, such as data on racial origin or sexual orientation, in the GDPR. This is why it may be currently unclear to what extent such processing is lawful in view of data protection legislation.

The proposed AIA may change this, as Article 10 (5) of the proposal explicitly mentions bias monitoring, detection and correction as a separate justification for the processing of sensitive categories of personal data as part of the quality standards for high-risk AI systems. If data controllers process sensitive data based on Article 9 (2) (g) GDPR in line with Article 10 (5) AIA proposal, they have to implement appropriate safeguards for the fundamental rights and freedoms of natural persons whose sensitive personal data are processed. The proposed law is still being negotiated at the time of writing this report.

The European Commission's ***Guidance note on the collection and use of equality data based on racial or ethnic origin*** highlights the relevance of collecting data on protected characteristics for the purpose of counteracting discrimination and inequality, with respect to racial or ethnic origin.

Endnotes

7 FRA (2019b), p. 166; and FRA (2020), p. 23.
8 Dutch Parliamentary Committee (2020), p. 14.
9 European Commission (2022).
10 FRA (2020).
11 STOA, European Parliamentary Research Service (2020), p. 15.
12 *Ibid.*, p. 16.
13 FRA (2018a); and FRA (2019a).
14 FRA (2018b).
15 European Parliament resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (**2016/2225(INI)**).
16 European Council (2019), p. 4
17 European Commission (2018).
18 Von der Leyen (2019), p. 13.
19 European Commission (2020).
20 European Commission (2021).
21 Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast) (Machinery Directive).
22 Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products (Product Liability Directive).
23 See Richardson *et al.* (2019).
24 FRA (2018c), pp. 97–102.
25 *Ibid.*, p. 231
26 Wachter *et al.* (2021a), pp. 24–25.
27 See, for example, Opinion of Mr Advocate General Léger, 31 May 1995, para. 58, in case CJEU, C-317/93, *Inge Nolte v Landesversicherungsanstalt Hannover*, 14 December 1995.
28 See CJEU, C-171/88, *Ingrid Rinner-Kühn v FWW Spezial-Gebäudereinigung GmbH & Co. KG*, 13 July 1989; CJEU, C-33/89, *Maria Kowalska v Freie und Hansestadt Hamburg*, 27 June 1990; CJEU, C-184/89, *Helga Nimz v Freie und Hansestadt Hamburg*, 7 February 1991; and CJEU, C-343/92, *M. A. De Weerd, née Roks, and others v Bestuur van de Bedrijfsvereniging voor de Gezondheid, Geestelijke en Maatschappelijke Belangen and others*, 24 February 1994.
29 FRA (2018c), pp. 230–232.
30 Wachter *et al.* (2021b).
31 See, for example, Musto *et al.* (2021); and Kennedy *et al.* (2020).
32 German Data Ethics Commission (2019), p. 23, point 53.
33 Wachter *et al.* (2021a), p. 6; Mittelstadt (2017); and Wachter (2019).
34 FRA (2020), p. 11.

2

FEEDBACK LOOPS: HOW ALGORITHMS CAN INFLUENCE ALGORITHMS

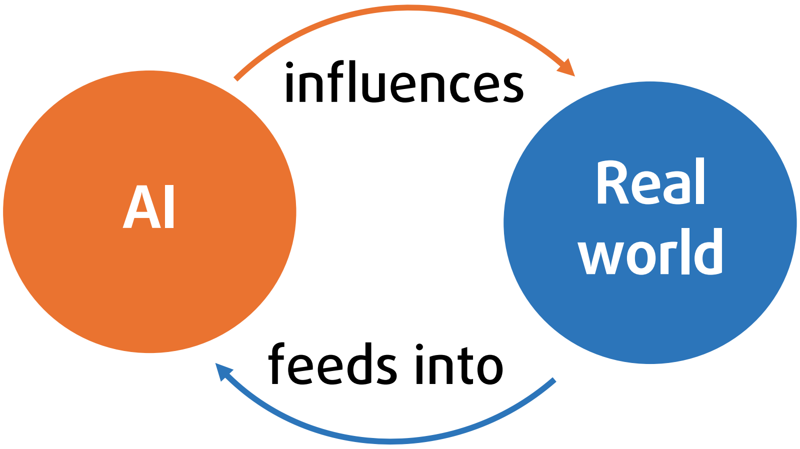
How could a feedback loop affect your day-to-day life?

Imagine living in a neighbourhood with a very strong police presence. While the standard of living and perceived safety in the neighbourhood increase considerably, police presence in the neighbourhood remains high and the police regularly stop and search people on the streets. Are the reasons behind the police presence based on correct and representative facts, and what role – if any – does the police's new AI-based crime prevention tool play in this? Does bias play a role?

This chapter provides an analysis of feedback loops, with respect to the potential for bias, based on a computer simulation to test a simplified version of a predictive policing algorithm in relation to crime occurrence in neighbourhoods. However, the aim of the analysis is not to examine predictive policing, but rather to analyse how feedback loops occur and under what circumstances.

A feedback loop occurs when predictions made by a system influence the data used to update the same system. Algorithms influence algorithms, because their recommendations and predictions influence the reality on the ground. For example, an algorithm's predictions of crime occurrence changes the behaviour of police officers, which in turn influences the detection of crime. The detected crimes are then fed back into the system. Feedback loops are common, and many machine learning systems have such built-in feedback. Figure 1 provides a simplistic illustration of a feedback loop.

FIGURE 1: SIMPLISTIC ILLUSTRATION OF A FEEDBACK LOOP



Source: FRA, 2022

Machine learning models include algorithms that have 'learned' to accomplish a given task by being 'trained' on data to identify a pattern that can be used to make predictions about new, unseen data. For example, credit-rating models can only determine a low-risk candidate based on who has been given credit in the past, an offensive speech detection algorithm can only detect offensive speech based on what has been tagged as offensive in the past and prediction models used in human resources contexts to screen potential applicants for a position can only update based on the actual future performance of approved applicants and not that of those applicants who were rejected.

Some machine learning models, such as so-called batch models or online learning models, continue to 'learn' after deployment. The model produces a prediction, decisions are made and the observed results are added to the training data for the next round of training. This is a classic feedback loop situation: predictions made by the system influence the data fed back into it for future predictions.[33]

In the policing context, predictive policing systems that can help determine which neighbourhoods should be patrolled are based on available existing crime data. Such data include observed and reported incidents. Feedback loops, as described above, can create a so-called self-fulfilling prophecy. For example, if the system 'detects' more crime in district A and decides to send more patrols there, more crimes will be recorded in district A, and the corresponding data will be fed back into the system, reinforcing the system's 'belief' that there is more crime in district A.

Feedback in a system is described as 'runaway' when it causes a 'winner takes all' situation – it ends up recommending only one solution and overestimates results. In predictive policing systems, this would mean that the system is likely to send police to the same neighbourhoods, based on the data fed into the system, regardless of the true crime rate. This would lead to overpolicing some neighbourhoods while underpolicing others.

The simulation study in this section will show, based on a simplified example, how several parameters may influence the formation of feedback loops. These include the following.

- **Crime reporting rates.** These include crimes reported to the police by witnesses or victims.
- **Police distribution.** This refers to the percentage of police patrols distributed across neighbourhoods.
- **Observability of crime.** This measures how likely it is that police detect a crime in a particular neighbourhood.

- **True crime distribution.** This is an assumed distribution of crime taking place, including reported, detected and unrecorded crime. Such a number is not known in reality but can be included in a simulation study, using a range of different values to explore its impact on feedback loop formation.

In addition, the influence of different algorithms is analysed, as are mitigation measures to limit the influence of feedback loops. The simulation is based on theoretical assumptions that are tested by artificially mirroring many iterations over time. While the results of the simulation exemplify how real-life situations may play out over time, it is not an analysis of real-world applications, which are likely to be more complex. However, in order to make as realistic assumptions as possible, some parameters are taken from real-world data, such as data on the general population's crime experiences and rates of reporting to the police, which are collected by crime victimisation surveys.

Section 2.1 provides an overview of the fundamental rights and applicable EU law relevant to the use case of predictive policing, the example used throughout this chapter, while Section 2.2 provides an overview of current actual use of predictive policing in the EU and Section 2.3 outlines known challenges linked to the use of predictive policing. Section 2.4 presents the results of the simulations carried out for this analysis, and Section 2.5 concludes.

2.1. EU LAW AND BIAS IN PREDICTIVE POLICING SYSTEMS

2.1.1. Fundamental rights affected by overpolicing and underpolicing

Overpolicing is understood, for the purpose of this report, as the disproportionate 'over'-presence of police in a particular area in relation to the true crime rate. Underpolicing means a disproportionate 'under'-presence of police in a particular area in relation to the true crime rate. Both can have adverse fundamental rights implications.[34] The debate concerning discriminatory and other adverse effects of predictive policing on fundamental rights has so far mostly focused on overpolicing and concerns that individuals present in overpoliced areas might be negatively affected, namely with respect to their fundamental rights.

Overpolicing can involve concrete action being taken against particular individuals, including police stops and searches, identity checks and intrusion in homes. These actions can affect the right to physical integrity (Article 3 of the Charter), the right to respect for private and family life (Article 7 of the Charter) and the right to data protection (Article 8 of the Charter). When individuals are arrested, the right to liberty and security (Article 6 of the Charter) is affected. Overpolicing may also have a 'chilling effect' on the way individuals express their views and/or gather in publicly accessible spaces, affecting the freedom of expression and information (Article 11 of the Charter) and the freedom of assembly and of association (Article 12 of the Charter).

While overpolicing may affect fundamental rights, including because of its potential link to racial profiling and similar discriminatory police activities, underpolicing can also be detrimental to fundamental rights. A lack of police presence in a particular area may put people living there at a higher risk of becoming victims of crime, posing a risk to a variety of fundamental rights, ranging from the rights to life and physical integrity (Articles 2 and 3 of the Charter) to the right to property (Article 17 of the Charter), for example. When individuals fear becoming victims of crime, this can also negatively affect their enjoyment of a range of fundamental rights.

Overpolicing may amount to discrimination where the negative effects associated with it (see above) lead to less favourable treatment of individuals based on protected characteristics. For instance, the use of predictive policing algorithms may result in overpolicing of areas mainly inhabited by certain ethnic minorities, whereby the area itself becomes a proxy for ethnic origin. The European Court of Human Rights (ECtHR) has consistently held that the state has a duty to investigate a potential causal link between police officers' alleged racist attitudes and mistreatment suffered by individuals at their hands.[35] The same critique should be true for biased algorithms.

If the initial assumption of the police that there will be more crime in district A than in district B is based directly on a protected characteristic (such as the dominance of a particular ethnic group in district A), this would amount to direct discrimination.[36] If, however, a place-based predictive policing model that pursues a seemingly objective goal (e.g. the prevention of crime) through seemingly objective means (e.g. a higher incidence rate of crime in a certain neighbourhood) ends up targeting groups associated with a protected characteristic more than others, this is indirect discrimination. This type of discrimination is more difficult to prove and is subject to justification if there is a legitimate aim and the measure is proportionate.[37]

As police resources are limited, overpolicing certain areas compared with others regardless of the true crime rate ultimately also affects the rights of individuals outside those overpoliced areas. Underpolicing, by failing to intervene when – for example – women report intimate partner violence, can also have significant fundamental rights consequences. The ECtHR has emphasised that states failing to adequately respond to certain forms of violence were considered in breach of Convention rights.[38] For example, the ECtHR found Turkey in breach of non-discrimination under Article 14 European Convention on Human Rights, on the basis of gender, in conjunction with the right to life and prohibition of ill-treatment under Articles 2 and 3. In that case, the police had failed to adequately address domestic violence in a case of clear threats by a known perpetrator of domestic violence.[39] There is extensive ECtHR jurisprudence clarifying that the state is obliged to carry out effective investigations when crimes occur.[40]

Police stops in the European Union

Around 14 % of the general population experienced a police stop during the past year and 27 % experienced one in the past five years, according to FRA's 2019 Fundamental Rights Survey, which is based on around 35,000 interviews across the EU, North Macedonia and the United Kingdom. Police stops more often concern men, young people, people from ethnic minorities, especially Muslims, and people who do not self-identify as heterosexual. For example, out of people who consider themselves to be part of an ethnic minority, 22 % in the EU-27 were stopped by the police in the 12 months before the survey, as opposed to 13 % of people who do not consider themselves to be part of an ethnic minority.

Immigrants' and ethnic minorities' trust in the police depends on how they experience being stopped by the police and whether or not they perceive the stops as ethnic profiling. The perception of being subjected to ethnic profiling when stopped by the police in the previous five years is most common among immigrants and descendants of immigrants from South Asia in Greece (89 %) and Roma in the Netherlands (86 %) and Portugal (84 %), according to FRA's Second European Union Minorities and Discrimination Survey (2016) and the Roma and Travellers Survey (2019). In comparison, other minorities surveyed had a lower perception of having experienced a discriminatory police stop. For example, the Russian minority in the Baltic states never felt that they were stopped by the police because of their ethnic background. In Poland and Slovenia, 9 % and 5 %, respectively, of recent immigrants from countries outside the EU felt they were stopped because of their ethnic origin.

Source: FRA (2021b).

In addition, the interaction of the police with the public in overpoliced or underpoliced areas can also be different. This is again based on assumptions or learned responses concerning these areas and the people who inhabit them. This may be reflected in the 'quality' of a police encounter with the public, for example with respect to a police stop, whereby those stopped in areas that are 'overpoliced' receive less respectful treatment than those stopped in areas that are 'underpoliced'. This is something that has been documented by several studies, but is – in this instance – beyond the scope of the current 'use case' model with respect to available data. See, for example, results of FRA surveys showing that certain groups, such as people with Roma and North African backgrounds, more often experience disrespectful behaviour by police.[41]

2.1.2. EU secondary law in relation to predictive policing

The Law Enforcement Directive[42] deals with the processing of personal data for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties. Article 11 Law Enforcement Directive addressed automated individual decision-making, which is prohibited unless authorised by an EU or Member State law that provides appropriate safeguards: "at least the right to obtain human intervention".[43] Moreover, authorities that wish to use predictive policing algorithms should carry out a data protection impact assessment.[44] However, the Law Enforcement Directive may not apply to place-based predictive policing (in contrast with person-based predictive policing), as predictions in the context of place-based predictive policing usually do not include personal data, but rather use aggregated statistics.

In addition, the decision-making may not be fully automated (i.e. based "solely on automated processing" and "without meaningful human involvement in the decision process") because the predictions support only decisions on where to send police.[45] While the simulations used for this analysis were fully automated, in the real world there is usually a form of meaningful human involvement when deciding where to allocate police patrols, which is why Article 11 Law Enforcement Directive would probably not apply to the real-world use of such algorithms.[46]

The AIA proposal (see Section 1.2) contains specific requirements for feedback loops pursuant to Article 15 (3) AIA, which addresses "AI systems that continue

to learn after being placed on the market or put into service". The design of such systems has to ensure that biased outputs due to feedback loops are "duly addressed with appropriate mitigation measures". This provision only applies to AI systems categorised as 'high risk'. This applies to law enforcement use cases in the area of predictive policing, but currently does not apply to place-based predictive policing systems. As the negotiations of the proposed regulation were ongoing at the time of writing this report, it is not analysed any further at this stage.

2.2. PREDICTIVE POLICING IN THE EUROPEAN UNION

Various law enforcement agencies across the EU have used or are currently using (or at least testing) algorithmic systems – so-called predictive policing systems – to predict when and where crime might take place.[47]

Predictive policing involves the application of analytical techniques – particularly quantitative techniques – to identify likely targets for police intervention and to prevent crime or solve past crimes by making statistical predictions.[48]

Beyond using crime statistics, the data such systems use for training and/or as input can include geographical data (such as landmarks and important infrastructure) and social, cultural and economic information. The algorithms

FRA handbook: *Preventing unlawful profiling today and in the future: A guide*

In developing and using algorithmic profiling, bias may be introduced at each step of the process. To avoid this and subsequent potential violations of fundamental rights, both information technology experts and officers interpreting the data should have a clear understanding of fundamental rights.

This FRA guide explains what profiling is, the legal framework that regulates it and why conducting profiling lawfully is both necessary to comply with fundamental rights and crucial for effective policing and border management.

For more information, see FRA (2018), **Preventing unlawful profiling today and in the future: A guide***, Luxembourg, Publications Office.*

used range from logistic regression[49] to more complex machine learning methods. However, for various reasons – trade secrets, data privacy issues, confidentiality of law enforcement activities for security purposes – very little is generally known about the algorithms used in predictive policing, the data used to train the models and the ensuing police actions taken as a result of the predictions. In summary, most predictive policing models are proprietary, making it difficult to research and understand their precise functioning. One exception to this is PredPol,[50] a tool used to predict high-risk areas of crime based on historical crime data, including information about the type of crime, its location and the time of occurrence.[51]

Predictive policing systems in Europe

Alongside internationally relevant predictive policing systems such as **PredPol** and **HunchLab**, other systems are currently being deployed in Europe. For example, one of the most widespread systems in Austria, Germany and Switzerland is **Precobs**, a German software developed by the Institut für musterbasierte Prognosetechnik.*

Precobs is a system for assessing the likelihood that certain areas will experience burglaries during a given time span.** The algorithm is based on the theory of near-repeat phenomena, which identifies burglaries that are likely to be followed by crimes in the vicinity.*** It uses geographical data, combined with police statistics on burglary locations, time of occurrence, items stolen and modus operandi, to deduce patterns corresponding to professional serial burglars and predict likely near-repeat burglaries.**** Here, theory and research pre-date the use of AI systems, notably in relation to criminological studies from the 1980s on crime 'hotspots'.*****

Two predictive policing systems were deployed in the Netherlands in 2019: SyRI and CAS. SyRI is a system designed to help the government identify individuals at risk of engaging in fraud in the areas of social security, tax and labour law. While PredPol, Precobs and CAS are examples of 'place-based predictive policing', SyRI provides an example of a 'person-based predictive policing' system. SyRI received a lot of media attention owing to a 2020 District Court of The Hague ruling (**ECLI:NL:RBDHA:2020:865**) that it violates the right to privacy as contained in Article 8 European Convention on Human Rights,****** and, as a result, was discontinued by the Dutch government. The second system, CAS, has been deployed on a national scale in the Netherlands, making the Netherlands the first country in the world to adopt predictive policing nationwide, and is still in use. Rather than focusing on individuals, CAS is programmed to identify future crime hotspots. Areas are divided into squares of 125 by 125 metres, which are categorised by the potential risk of crime occurring in this area. Colour coding these squares results in so-called heat maps, which are used by the police to investigate areas or distribute police patrols.*******

* *Lattacher, S. (2017), '**Predictive Policing: Frühwarnsystem für die Polizei**', Magazin öffentliche Sicherheit, Vol. 3/4, pp. 11–12; and Egbert, S. and Krasmann, S. (2019), '**Predictive Policing. Eine ethnographische Studie neuer Technologien zur Vorhersage von Straftaten und ihre Folgen für die polizeiliche Praxis**', Projektabschlussbericht, Hamburg, Hamburg University, 30 April 2019.*

** *Gerstner, D. (2018), 'Predictive policing in the context of residential burglary: An empirical illustration on the basis of a pilot project in Baden-Württemberg, German', European Journal for Security Research, Vol. 3, pp. 115–138.*

*** *Townsley, M., Homel, R. and Chaseling, J. (2003), 'Infectious burglaries. A test of the near repeat hypothesis', British Journal of Criminology, Vol. 43, pp. 615–633.*

**** *Egbert, S. and Krasmann, S. (2019), '**Predictive Policing. Eine ethnographische Studie neuer Technologien zur Vorhersage von Straftaten und ihre Folgen für die polizeiliche Praxis**', Projektabschlussbericht, Hamburg, Hamburg University, 30 April 2019.*

***** *Sherman, L., Gartin, P. and Buerger, M. (1989), 'Hot spots of predatory crime: Routine activities and the criminology of place', Criminology, Vol. 27, No. 1, pp. 27–56.*

****** *Strikwerda, L. (2020), '**Predictive policing: The risks associated with risk assessment**', Police Journal: Theory, Practice and Principles, 6 August 2020, pp. 1–15.*

******* *Ibid.*

2.3. SOME CHALLENGES LINKED TO THE USE OF PREDICTIVE POLICING

The literature on predictive policing currently raises three main problems with predictive policing systems:

- lack of transparency and accountability[52]
- biased data[53]
- the potential for runaway feedback loops.[54]

The **lack of transparency and accountability** covers two aspects. First, owing to intellectual property law, and often also the complexity of the algorithms used, law enforcement officers themselves may not have the information or training required to fully comprehend what an algorithm is doing and why. This makes it difficult to detect errors or biases in model predictions. In addition, the lack of transparency restricts the possibility for independent research, which results in a lack of objective assessment of the mechanisms of action and effectiveness of the predictive policing systems.

The second issue is **biased data**. FRA previously pointed to the recurrent concern that the reliance on historical crime data – which may be biased or incomplete – could lead predictive policing systems to reproduce and entrench existing discriminatory practices.[55]

The main objective of predictive policing is to identify areas at high risk of crime (or individuals at high risk of committing crimes) so that police can take action to prevent such crime. These systems are based on the premise that past crime events contain the patterns for predicting future crime events, and therefore require historical crime data to function. However, decades of criminological research have shown the limitations of such an approach, as police databases are not a complete census of all criminal offences and do not constitute a representative random sample.[56]

To mitigate the limitations of official crime statistics in accurately assessing the 'real' extent and nature of crime, a number of countries carry out victimisation surveys, which randomly sample the population and ask them about their experiences of crime, ranging from property crime through to violent crime, and more recently online fraud and related internet crime. Importantly, these surveys ask people whether they report their experiences of crime to the police, which allows for an estimate of how much crime is 'undercounted' in official crime statistics – the so-called dark figures. For example, the National Crime Victimization Survey has been carried out in the United States since 1973, the British Crime Survey has been undertaken in the United Kingdom since 1982 and the Swedish Crime Survey has been carried out in Sweden since 2006.[57] FRA's surveys adopt a classic crime victimisation survey model when asking about experiences of crime victimisation and reporting to the police, ranging from targeted surveys of specific groups in the population (e.g. ethnic minorities and immigrants, Roma and Jewish respondents) or specific subject areas (violence against women), alongside the FRA Fundamental Rights Survey of the general population.

The third challenge associated with predictive policing systems is the subject of the present analysis: the potential for runaway feedback loops.

Police crime statistics

Police crime statistics consist of two kinds of records/ sources (each of which may be associated with bias and errors):

crime incidents that are observed or detected by the police themselves (referred to here as 'detected crime');

crime incidents that are reported to the police by victims, witnesses or third parties.[*]

For 'detected/observed crime', it should be noted that not all crimes are equally observable – for example violence and drug-related offences committed openly in public places versus the less visible crimes of business fraud and tax evasion[**] – causing certain types of crime to be overrepresented in crime statistics. Furthermore, police practices themselves can be biased or prejudiced, meaning that what gets perceived as a crime[***] or who is stopped and controlled[****] (and hence more likely to be detected) largely depends on demographics. In turn, not all crime incidents that are reported to the police are recorded, and therefore may not feature in official police crime statistics.

At the same time, victimisation surveys also provide evidence that crime reporting varies – by type of crime and by socio-economic status of victims.[*****] The combined effect for the police crime data is that crime observability and overall detection can depend on crime type and on demographic characteristics of the perpetrator and/or the victim.

[*] *Dreißigacker, A. (2017),* Befragung zur Sicherheit und Kriminalität: Kernbefunde der Dunkelfeldstudie 2017 des Landeskriminalamtes Schleswig-Holstein, *Hannover, Kriminologisches Forschungsinstitut Niedersachsen; and Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. (2018), '**Runaway feedback loops in predictive policing**',* Proceedings of the 1st Conference on Fairness, Accountability and Transparency, *PMLR 81, pp. 160–171.*

[**] *FRA (2020),* **Getting the future right – Artificial intelligence and fundamental rights**, *Luxembourg, Publications Office.*

[***] *Richardson, R., Schultz, J. and Crawford, K. (2019), '**Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice**',* NYU Law Review, *Vol. 94, No. 192, p. 218.*

[****] *FRA (2021),* **Your rights matter: Police stops - Fundamental Rights Survey**, *Luxembourg, Publications Office, 25 May 2021; and Lum, K. and Isaac, W. (2016) 'To predict and serve?',* Significance Magazine, *Vol. 13, No. 5, pp. 14–19.*

[*****] *Murrià, M., Sobrino, C. and González, C. (2020), 30 años de la Encuesta de Victimización del Área* Metropolitana de Barcelona. Vigencia y uso de las encuestas de seguridad en las metrópolis, *Barcelona, Institut d'Estudis Regionals i Metropolitans de Barcelona.*

2.4. SIMULATING FEEDBACK LOOPS IN POLICING

Two main simulations were conducted in order to demonstrate how feedback loops may occur in predictive policing.

- The first simulation is built on a simple machine learning classification model. It explores which parameters influence the process of feedback loop formation generally.
- The second simulation employs a more complex model, offering a more realistic setting in which feedback loops form in predictive policing systems.

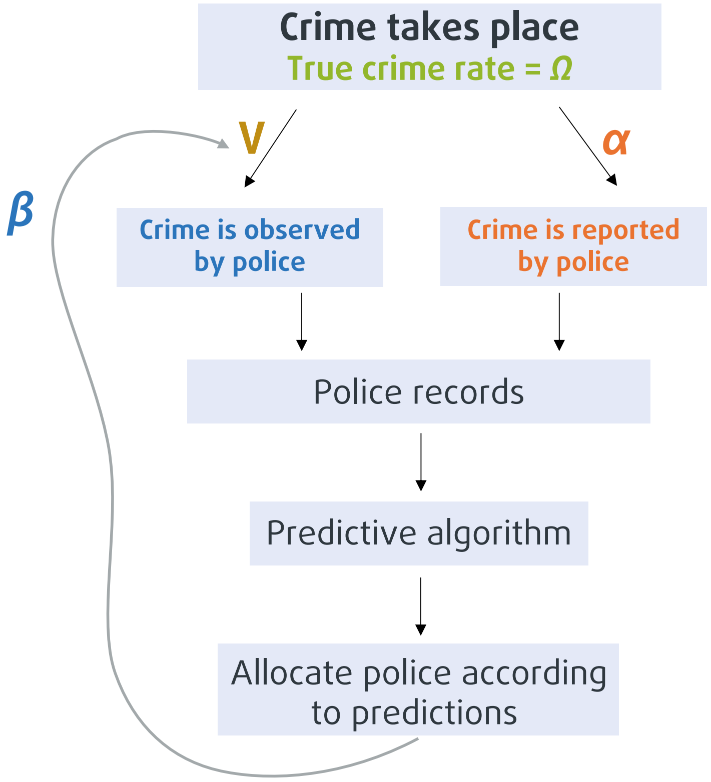
'Synthetic' datasets were generated for the simulations, meaning that data were artificially produced and randomly created by a computer to have the necessary statistical properties. It can be understood as the simulation of a distribution of crime in a fictitious city. These datasets do not correspond to real crime events. However, some parameters are based on real-world data.

As mentioned above and depicted in Figure 2, the simulation included several parameters that were investigated for their influence on the formation of feedback loops. These include the crime reporting rates (*α*), police distribution (*β*), observability of crime (*V*), which is the likelihood of detecting a certain crime) and the assumed 'true' crime distribution (*Ω*).

Parameters *α* and *V* take into account all recorded crimes (reported and detected). The true crime distribution additionally includes (assumed) non-recorded crime. Parameters *α*, *V* and *Ω* remain constant throughout the simulation; *β* is initially determined, and is the only value that changes over time, identifying the formation of feedback loops.

These are basic assumptions of the simulation study, which is limited with respect to the number of parameters investigated and analysed.

FIGURE 2: SCHEMATIC DESCRIPTION OF THE SIMULATED PREDICTIVE POLICING PROCESS


Crime takes place
True crime rate = Ω
V
α
β
Crime is observed by police
Crime is reported by police
Police records
Predictive algorithm
Allocate police according to predictions

Source: FRA, 2022

Two simulations were conducted, which vary according to:

- the policing strategy (i.e. the allocation of police to areas based on the predictions), including allocation of police according to levels of crime (**effective policing**) and allocation of the majority of police to a fixed number of most affected areas (**hotspot policing**);
- the prediction algorithms used, including simple probabilistic models and more complex machine learning models;
- the number of areas (also referred to as neighbourhoods or cells), which differs as only two are used in the first simulation whereas multiple areas are used in the second simulation.

Further assumptions include the following:

- although the average number of true crime events is kept constant within each simulation, it is assumed that only one third of the true number of crime events are observed and recorded;
- a police visit to a cell increases the likelihood that a crime event is detected by a factor of five. This is only an assumption, and is used to hold this parameter constant in order to ease interpretation of the model's results. In reality, the likelihood that crime is detected may vary across neighbourhoods based on different police behaviour within these neighbourhoods. FRA's surveys show differences across ethnic groups' experiences with police treatment (see Section 2.1.1 above).

The simulation is necessarily limited to a smaller number of assumptions in order to keep various factors constant and allow for testing of variations among the parameters selected. The details of the assumptions and parameters are outlined in Annex II of this report.

Mitigation strategies were employed and tested as the final step. Mitigation strategies were suggested based on the sources of bias identified in the research; they were then tested in both simulation experiments. Such mitigation strategies include behavioural changes of actors such as victims or the police, and technical methods to remove data imbalances. This step includes assuming that reporting rates are changing or varying on the one hand, and that technical solutions to avoid predictions too closely follow the training data (i.e. technical measures to avoid overfitting) on the other. Further explanations of mitigation strategies and the results of their testing are provided in Section 2.4.3.

2.4.1. Simulation 1: Exploring different sources of bias

The first results are based on a simple probabilistic model in only two neighbourhoods. This means the police patrols are distributed according to the crime distribution based on 'historical' police data. In this simulation, the police are always sent in proportion to the crimes observed, and no statistical or machine learning tools are used. In this case, no runaway feedback loops are formed when all the following conditions are satisfied.

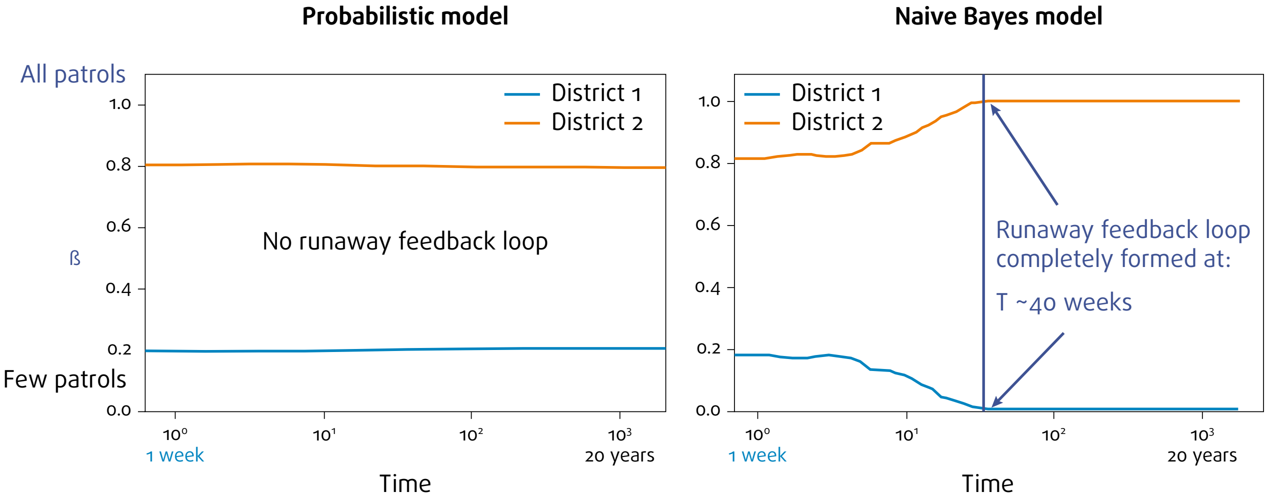
- The **true crime distribution** (Ω) is uniform, which means that crime can happen in every neighbourhood with the same likelihood.
- **Crime reporting rates** (α) are the same for all neighbourhoods. In other words, the proportion of victims reporting crime behaviour is the same across neighbourhoods.
- The police act in exactly the same way when sent to different neighbourhoods.

This is true for any value of initial patrol distribution, and reproduces previously published theoretical predictions.[58] However, as soon as the true crime distribution deviates from being uniform, runaway feedback loops are formed, if assuming the crime reporting rates are small (around 0.1) or zero and no mitigation measures are adopted.

Only when a machine learning algorithm is used may a feedback loop occur. A simple probabilistic model and a simple machine learning model called naive Bayes are used in this example. This is shown in Figure 3. The true crime rate was the same in both neighbourhoods, but the initial allocation of police was 20 % in one neighbourhood and 80 % in the other. It can be observed that, for the probabilistic model, the distribution of patrols *β* (represented by the orange and blue lines) did not change over time (the lines remained horizontal). Despite the true distribution of *Ω* being uniform, the initial historical bias was maintained throughout the simulation. In contrast, the naive Bayes predictions with the same parameters gradually contributed to the generation of a runaway feedback loop that assigned, after 40 weeks, 100 % of the police resources to district 2. Even in the situation in which the true crime rate is uniform and reporting rates to the police are zero in both neighbourhoods, the naive Bayes model forms a runaway feedback loop. This is the first of several instances in which the introduction of a machine learning model increases the unpredictability of the predictive policing system.

Even when the true crime rate and the corresponding allocation of patrols are both uniform (i.e. the same or constant), the naive Bayes model ends up with a runaway feedback loop. However, where the model ends up assigning 100 % of the police patrols is entirely a matter of chance: when running the same simulation 10 times, the model assigned all patrols to district 1 on five occasions and it assigned all patrols to district 2 on the other five occasions. The amount of time it took for the runaway feedback loop to form also varied greatly: it took 75 weeks in some cases and even more than 400 weeks in one case. This means that the use of machine learning algorithms can exacerbate the formation of a feedback loop.

FIGURE 3: RESULTS OF SIMULATION 1: ALLOCATION OF PATROLS OVER TIME – SIMPLE PROBABILITIES VS SIMPLE MACHINE LEARNING ALGORITHM



Source: FRA, 2022

Thus, we see that as soon as a predictive model is included in the system, runaway feedback loops are formed for all combinations of parameters. Machine learning models can amplify small differences, which may be enshrined in historical data: the simulations end up assigning all the police patrols to the neighbourhood with the highest crime rate in the input data. The determination of where the runaway feedback loop will assign all the police patrols is random (i.e. by chance) if the differences in true crime rates are small (1 % or less). Moreover, this behaviour is reproduced also by logistic regression. Importantly, such behaviour is observed if no additional measures are taken, whereas in real-life contexts humans would intervene to interpret, question and apply the results. These initial simulations simply show that machine learning algorithms tend to react to random signals/patterns and may exacerbate them.

In addition, various other sets of values of the relevant parameters were explored, and their influence individually and in several combinations was investigated. The execution of several simulations highlights different possible sources of bias in the system.

Not surprisingly, the level of **crime reporting** influences feedback loops. When crime reporting rates are greater than zero and equal for all neighbourhoods, the formation of runaway feedback loops is reduced. However, if a neighbourhood with a low crime reporting rate coincides with a neighbourhood where the true crime rate is higher, this mitigates the formation of feedback loops, equilibrating reported and observed/detected crime. However, if the place where victims report more crime coincides with where the true crime rates are higher, then the mitigation of feedback loops is diminished.

The opposite case becomes relevant when the difference in crime reporting rates between neighbourhoods is large (more than doubled) and the true crime distribution is close to uniform. In that case, the neighbourhood with lower 'true crime' ends up with the largest portion of police patrols. This phenomenon enhances the relevance of crime reporting, because, as has been observed, feedback loops usually form around the neighbourhood with the highest 'true crime'. However, here it can be observed that the 'true crime rates' can be distorted by excessive differences in the crime reporting behaviour across neighbourhoods.

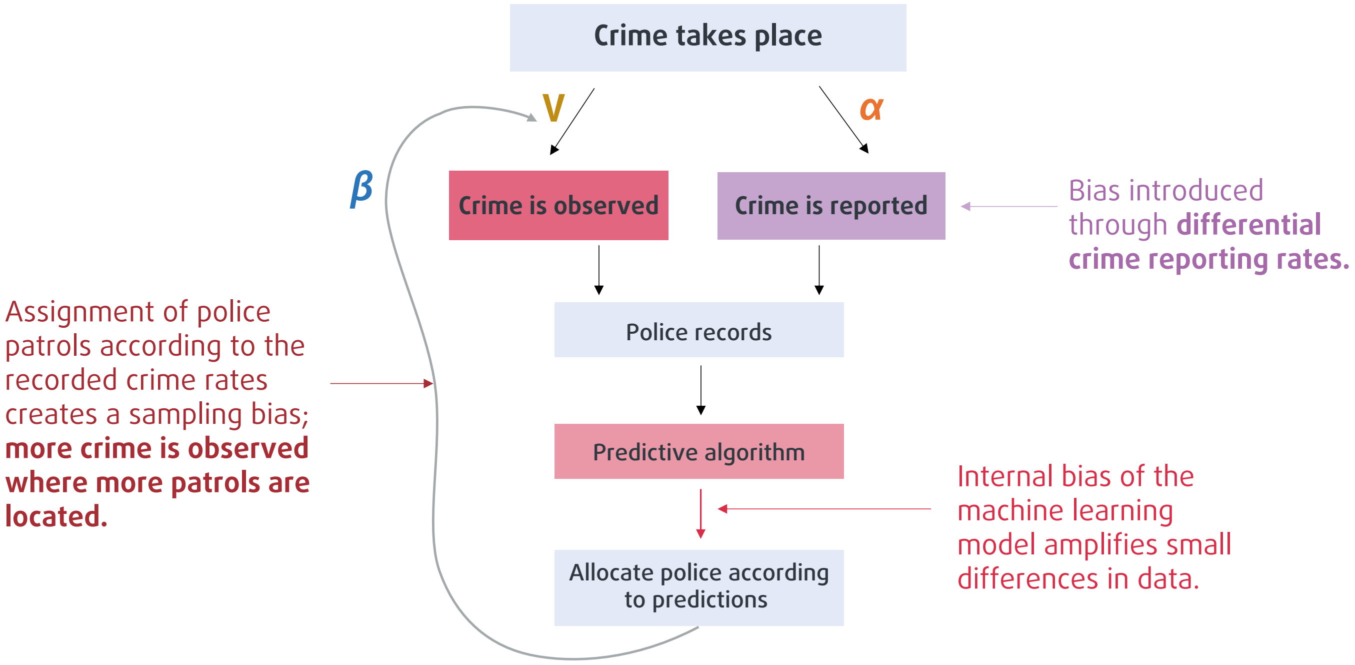
Finally, we also consider the parameter **crime observability**. This indicates the likelihood of a crime being observed if police are present in a neighbourhood. When crime reporting is zero, crime observability may be more relevant than the true crime rate, and the runaway feedback loop ends up sending all police patrols to an area where there is a lower overall crime rate but where crime is more observable. This effect can sometimes be mitigated by crime reporting rates under certain circumstances, but it can also be reinforced by them. The result does not change much if the initial crime distribution is varied.

Three different sources of bias were analysed in the first experiment (these are summarised below and visualised in Figure 4):

- a first source of bias can be observed when crime is reported through differential crime reporting rates;
- a second source of bias may stem from machine learning models that may overreact to random noise (i.e. by chance) in the data and amplify small differences;
- a third source of bias is related to differential crime observability.

Other sources of bias were not included in this simulation. For example, it may be assumed that the police apply different policing methods and have a different rapport with the inhabitants of an area depending on whether they regard an area as having high or low levels of crime. While the simulation cannot cover all potential sources of bias, the effect of increasing bias through feedback loops holds true.

FIGURE 4: SOURCE OF BIAS IN SIMPLIFIED POLICING ALGORITHMS


Crime takes place
V
α
β
Crime is observed
Crime is reported
Bias introduced through **differential crime reporting rates.**
Assignment of police patrols according to the recorded crime rates creates a sampling bias; **more crime is observed where more patrols are located.**
Police records
Predictive algorithm
Internal bias of the machine learning model amplifies small differences in data.
Allocate police according to predictions

Source: FRA, 2022

Based on these findings, the following conclusions can be drawn.

- The initial distribution of patrols seems to be the least relevant factor for the formation of feedback loops compared with the other relevant parameters.
- The internal bias of machine learning models accelerates the formation of feedback loops; even in the situations in which no real differences could be amplified, it captures random variation, creates fictitious differences and progressively amplifies them.
- When crime reporting rates are equal across neighbourhoods, they play an important role in mitigating the formation of feedback loops; the extent of the mitigation depends on the true crime rates. However, erroneous allocation of police patrols may occur when crime reporting rates differ significantly across neighbourhoods but the true distribution of crime is close to uniform.
- Erroneous allocation of police patrols may also occur when crime observability differs across neighbourhoods.

Overall, based on the fictitious simulation cases, we found that crime reporting rates (α), crime observability (V) and the true crime distribution (Ω) are relevant for the formation of feedback loops. In addition, we identify three types of sources for potential misallocation of police patrols.

The first, due to runaway feedback loops, occurs when police are excessively allocated in the neighbourhood with the highest crime rate.

The second occurs when the neighbourhood with a lower crime rate has a higher concentration of more observable crime: feedback loops form, sending police patrols to the neighbourhood with less crime (because it is more observable).

The third occurs when crime reporting rates vary greatly (20 % or more) between neighbourhoods: the assignment of more police to neighbourhoods with less crime but more reporting clearly leads to erroneous allocation of patrols.

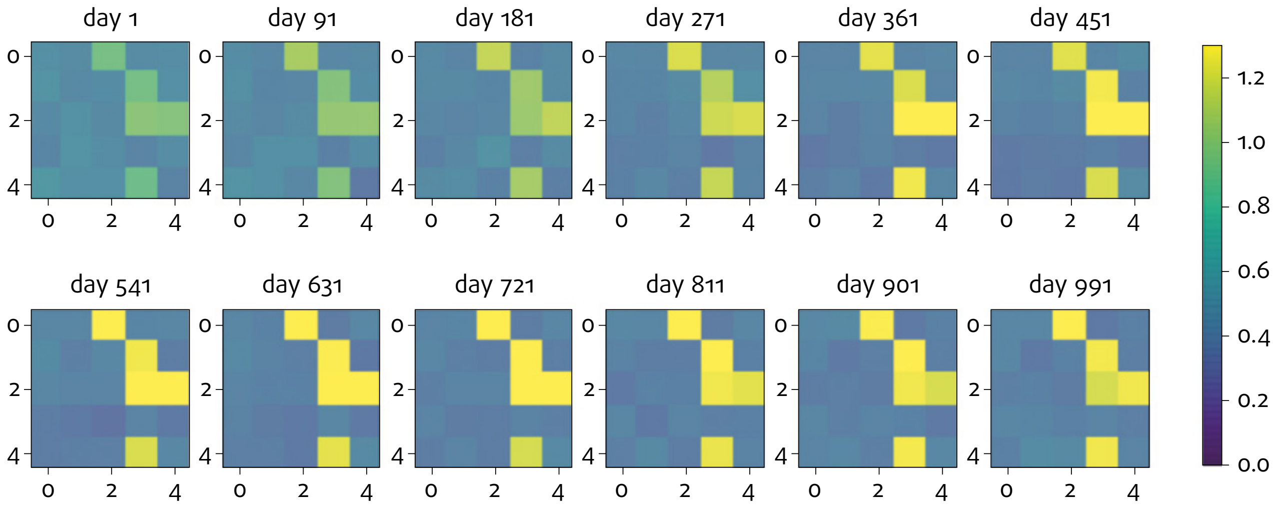
A recent study analysing victimisation datasets in Bogotá, Colombia, draws similar conclusions about the bias introduced by differential crime reporting rates.[59]

2.4.2. Simulation 2: Earthquake policing model and hotspot policing

The earthquake policing model (see Annex II for a description) has a different set-up from the first experiment. Its spatial distribution is different, as there are more neighbourhoods, and the method for allocating police patrols is different – it is based on hotspot policing.

The simulation starts by observing how feedback loops are formed in 25 neighbourhoods when the true crime rates and the crime observability are the same across all the neighbourhoods, but the historical crime rates differ, with five neighbourhoods having higher rates than the other neighbourhoods. Figure 5 shows that a feedback loop is formed around the five top cells. After 365 iterations of the simulation (i.e. one year, as one iteration represents one day), the estimated rates of the five cells with the highest historical crime rates increase whereas those of the other cells reduce, despite the true crime rate remaining constant and uniform. During the second year, the distribution is maintained except for some random fluctuations in the estimated values.

FIGURE 5: PREDICTED CRIMES PER DAY, WITH DIFFERING HISTORICAL CRIME RATES PER NEIGHBOURHOOD



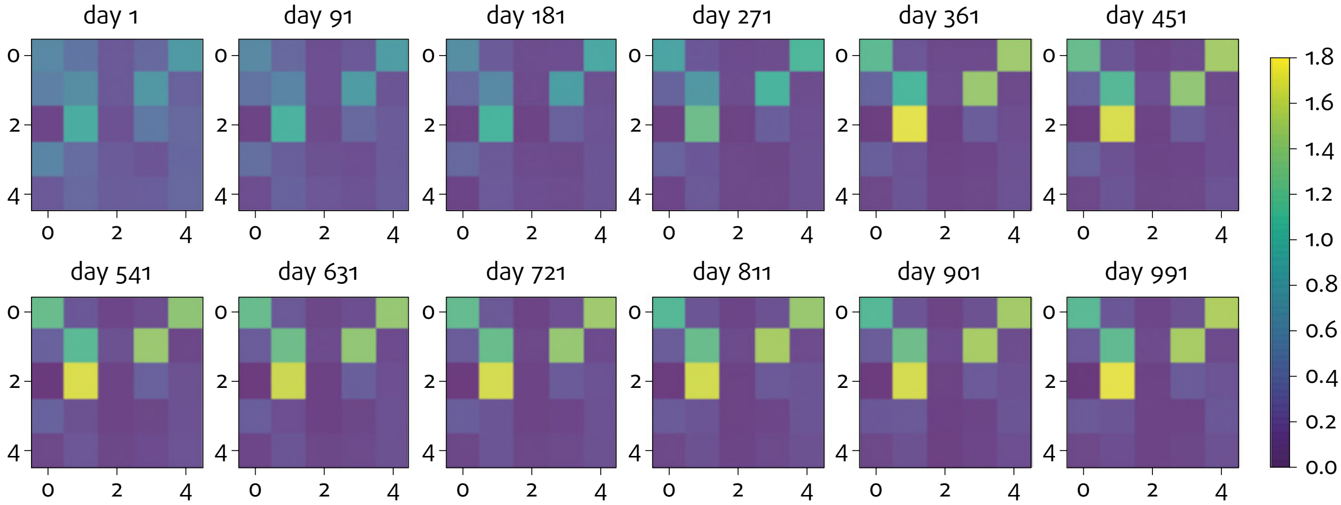
Notes: The cells correspond to areas in a fictitious city, and the colours indicate the predicted number of crimes in each cell. These are not necessarily whole numbers (integers), as the predictions are based on averages.

Source: FRA, 2022

Historical crime rates are also relevant if the true crime rates are, to some extent, higher in other neighbourhoods, but this depends on the magnitude of the difference of the rates. This result is not surprising and, in fact, also reflects an unlikely situation. The historical crime rates are to some extent linked to actual crime rates. However, it is still important to acknowledge that historical data drive the algorithm's predictions and future crime observations.

In the next example, unbiased data are assumed, which means that the historical data are equal to the true crime rates. Here, we can see that the patterns are strongly reinforced and overestimated after one year. The process is shown in Figure 6. The same simulation was also observed in another simulation of 138 cells. An overestimation in the 20 cells with the highest historical crime rates could be observed after one year (i.e. 365 iterations).

FIGURE 6: PREDICTED CRIMES PER DAY, WITH HISTORICAL CRIME RATES REFLECTING TRUE CRIME RATES



Notes: The cells correspond to areas in a fictitious city, and the colours indicate the predicted number of crimes in each cell. These are not necessarily whole numbers (integers), as the predictions are based on averages.

Source: *FRA, 2022*

Without mitigation techniques, after 365 iterations (i.e. one year) of the simulation in all tested conditions, feedback loops are formed. This means that the estimated rates in the cells with the highest historical crime rates (a fixed number, n, of 'top cells') have significantly increased, reducing the estimated values for the other cells. Furthermore, even when the historical data are unbiased, the set-up of the algorithm will form a feedback loop concentrated on the *n* top cells and distort the estimation of the true crime rate.

It is important to pay attention to the results of simulation 2, since the historical rates completely dominate the final result. The large difference in initial crime rates (the top cells have 50 % more crime than the other cells), compared with a true crime distribution close to uniform, leads to the preservation of the initial historical crime rates, and to the overestimation of crime in the top five cells.

2.4.3. Bias mitigation strategies

The formations of feedback loops, as described above, are based on theoretical models run repeatedly on a computer. This is a computer simulation creating artificial data (although to some extent initially based on real data, as described in Annex II). In reality, policing is much more complex and police consider many more aspects before sending patrols into certain neighbourhoods. Police work with people and are present on the streets, which is why a purely computer-based simple simulation cannot reflect reality and all its complexities. The above simulations exemplify some sources of bias that may lead to overpolicing and underpolicing of certain neighbourhoods. This may put people at a disadvantage if certain neighbourhoods are composed of people from particularly disadvantaged groups, most notably ethnically segregated neighbourhoods.

The following aspects can mitigate feedback loop formation.

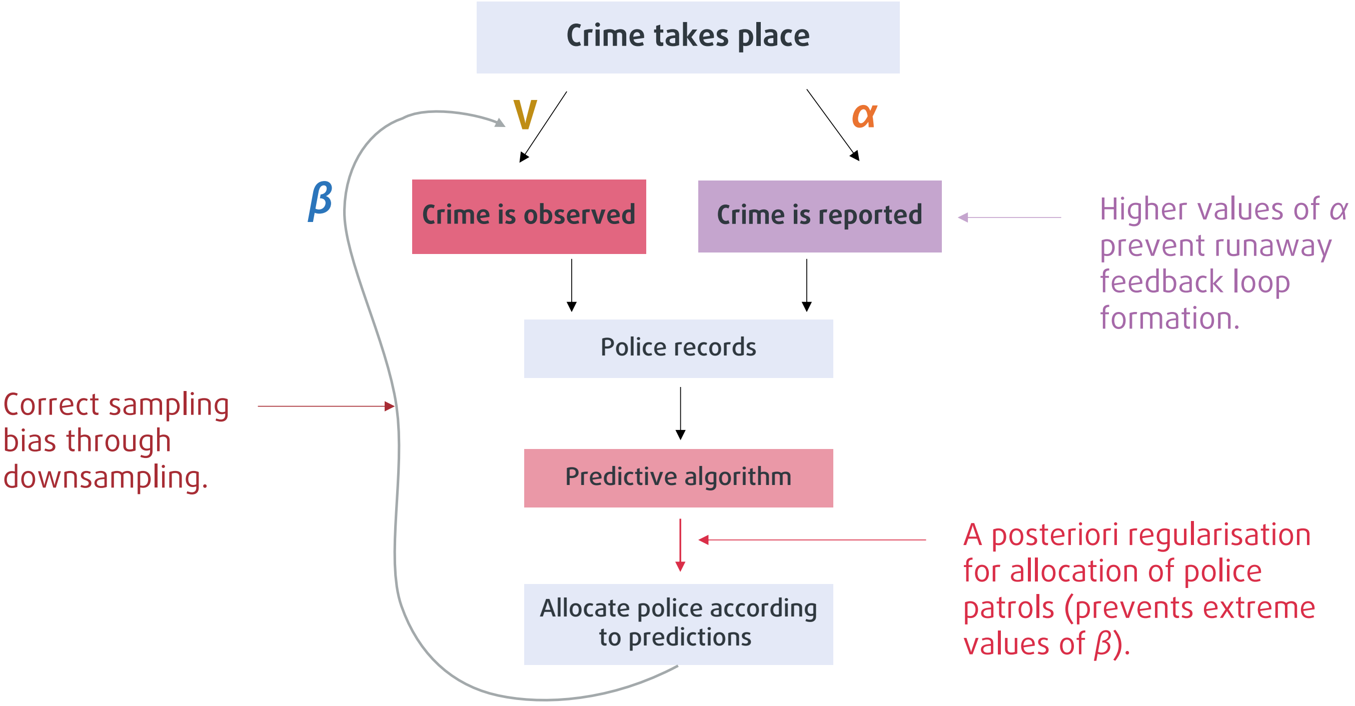
- As one of the sources of bias is low or biased crime reporting rates, increasing reporting rates can reduce feedback loops that may lead to overpolicing.
- Technical solutions to prevent values from becoming too extreme are needed. Machine learning is known to potentially focus too much on patterns in training data. This is called 'overfitting'. To prevent algorithms from predicting values that are too extreme, technical solutions called

regularisation have to be employed. This means that extreme predictions are avoided by adding a mathematical restriction to the algorithm. The choice of the value of such restriction needs close scrutiny to prevent feedback loops and, at the same time, allow predictions to be useful.

- Other solutions have been suggested in the literature, such as 'downsampling', which was introduced by Ensign *et al.*[60] This is based on assigning certain probabilities of recording to crime events to counteract the overly strong predictions.

These mitigation techniques are placed in the prediction cycle in Figure 7. While each technique has its limitations, a combination of mitigation techniques offers the most promising approach to mitigating bias.

FIGURE 7: BIAS MITIGATION TECHNIQUES REDUCING FEEDBACK LOOP FORMATION


Crime takes place
V
α
β
Crime is observed
Crime is reported
Higher values of α prevent runaway feedback loop formation.
Police records
Correct sampling bias through downsampling.
Predictive algorithm
A posteriori regularisation for allocation of police patrols (prevents extreme values of β).
Allocate police according to predictions

Source: FRA, 2022

2.5. ADDRESSING FEEDBACK LOOPS: CONCLUDING POINTS

The simulation experiments in this report were simple models that used only automated decision-making based on a limited set of criteria and available data. However, a real-life predictive policing system would be embedded in law enforcement agencies, with humans interacting and making decisions at all stages. A more promising approach to feedback loop mitigation would benefit from a case-specific investigation into the entire predictive policing life cycle, and the installation of context-relevant review and control procedures. Taking this into account, a real-life context-specific model of potential police bias – rather than the simulation model looked at here – would need to take into account sources of bias with respect to different actors that have an impact on a system. The following should be considered in such a model.

- As reporting rates affect policing, it is the victims of and witnesses to crime in the general population who are relevant in mitigating overpolicing and underpolicing. Feedback loops can thus be mitigated by increasing awareness in society and trying to improve crime reporting rates in general, particularly in neighbourhoods where reporting rates are low and linked to vulnerable groups in society. Increasing trust in the police is one important way of enhancing reporting rates, whereby the onus should not only be on the public to report crime, but also on the police to encourage crime reporting and make it accessible, through building trust. FRA's Fundamental Rights Survey showed that the lack of trust in the police is one reason for people not reporting burglary, either because of a general lack of trust (7 %) or because they expect the police would not do anything about it (25 %). Moreover, those who experienced police stops based on assumed ethnic profiling trust the police less.[61]

- The police play an important role when it comes to different rates of detecting crime, which – among other considerations – is related to the 'observability' of various crimes (such as street crime versus fraud). Fundamental rights-compliant policing should address police behaviour that is potentially different depending on which neighbourhoods police are patrolling. Awareness raising regarding policing behaviour that leads to differential crime observability supports fairer policing in this respect. Police may see prejudices confirmed through algorithms, and, more generally, may put too much trust in predictions produced by algorithms. Such an overreliance can hamper the necessary human review of algorithmic outputs.[62] Awareness about limitations and possible failures of predictive policing algorithms among police officers is one crucial aspect to avoid unfair policing practices as a result of police overreliance on algorithmic predictions.
- AI developers can play an important role when it comes to addressing the internal bias of machine learning algorithms. Police and developers, working together, can critically examine the assignment of police patrols according to the observed crime rates. The application of technical mitigation techniques, based on simulation and testing prior to deployment of systems, is required to avoid inefficient and potentially discriminatory policing practices.
- Affected communities can also be involved in talking to AI developers and police authorities, to better understand potential bias.

These results show that feedback loops can easily occur in fully automated settings. While strategies with a single focus, such as debiasing historical crime data, do not seem to be very effective, simple technical solutions to bias mitigation (such as regularisation methods) showed some success. The simulations also indicate that the main danger of runaway feedback loops has systemic causes. These are, in particular, differential crime observability and crime reporting rates. Left unaddressed, these issues have the potential to perpetuate bias and discrimination. A deeper investigation into these sources of bias is needed before predictive policing systems can be safe to deploy.

This chapter looked at simulations of simplified predictive policing algorithms. However, the same kinds of machine learning models are used in other predictive situations – employment decisions, credit rating, fraud prediction, to name a few – and so the findings from this section can have wider implications for the use of AI and its impact on fundamental rights in various contexts. For example, a credit risk model can only learn about future outcomes of credit repayment from those applicants who were granted credit and hence were considered low risk. High-risk applicants will probably not receive a loan, and so there is no additional information as to whether this decision was correct or not.

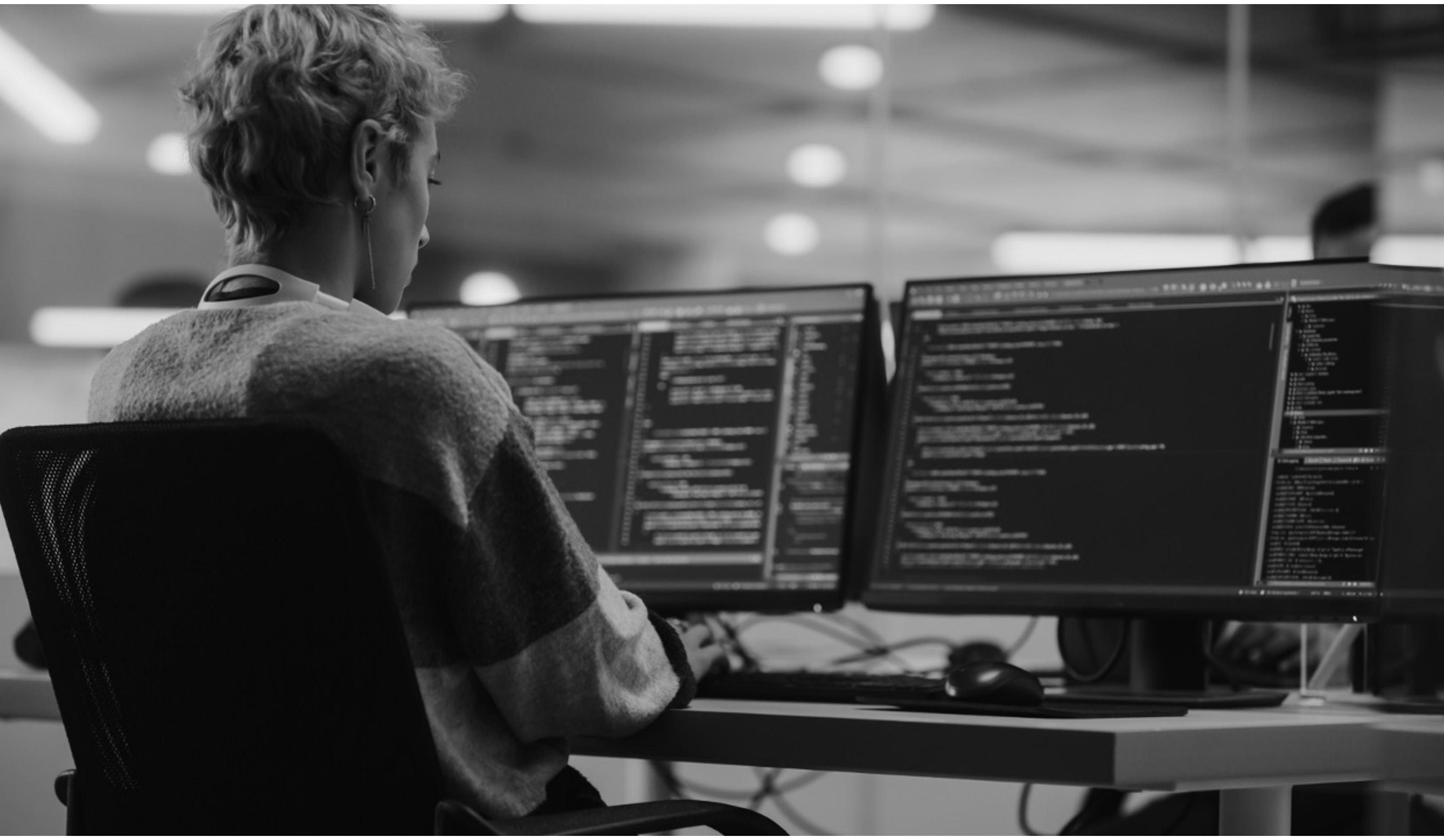
Such models also run the risk of runaway feedback loops. Many of the lessons learned in this simulation also apply. These include the necessity of maintaining a 'clean' source of fresh data, meaning that the data are not influenced by model predictions, as in the case of crime reporting for predictive policing, the importance of how decisions based on model predictions are made, and the importance of effective control and review throughout the model's life cycle.

Challenges encountered when researching feedback loops

This report shows results of a simulation of predictive policing to see how predictions evolve over many iterations. Several challenges are linked to researching feedback loops. The first challenge relates to availability of data on crimes, including information on the location, time and type of crime. Such datasets are not readily available in EU Member States (they are more readily available in the United Kingdom and the United States). In general, the granularity of available crime data in the public domain is limited, which also reduces the ability to research and simulate real crime events over space and time.

There is limited information available on the implementation of commercially developed algorithms that are used by law enforcement. Without such information, the potential biases, the feedback loop formation and the mitigation strategies are difficult to study accurately. The experiments described above show that certain techniques and methods can mitigate the development of feedback loops, while training and guidelines for police officers could potentially mitigate the effects of existing feedback loops. However, transparency is lacking because of trade secrets and the fact that these algorithms are owned by private companies and are thus proprietary. Currently, researchers and other relevant actors cannot test the exact functionality of proprietary algorithms and the extent of their effects, due to either a lack of information on the algorithm or a lack of available data.

The simulation experiments in this use case used simple models that used only automated decision-making. However, a real-life predictive policing system would be embedded in a law enforcement agency, with humans interacting and making decisions at all stages. In turn, human intervention should also be analysed in relation to how data based on predictive policing models are used and interpreted with respect to potential bias.



Endnotes

33 Ensign *et al*. (2018).
34 See FRA (2021b).
35 See, for example, ECtHR, *Boacă and Others v. Romania*, No. 40355/11, 12 January 2016, para. 108.
36 See FRA (2018d), p. 23.
37 See *Ibid*., pp. 23–24.
38 ECtHR, *Osman v. the United Kingdom*, Merits, No. 23452/94, 28 October 1998, paras 115–116: "Article 2 of the Convention may also imply in certain well-defined circumstances a positive obligation on the authorities to take preventive operational measures to protect an individual whose life is at risk from the criminal acts of another individual."; ECtHR, *Đorđević v. Croatia*, Merits, 24 July 2012, No. 41526/10, paras 138–139: failure to protect a disabled person from harassment is judged a violation of the prohibition of ill-treatment under Art. 3.
39 See, for example, in the context of domestic violence, ECtHR, *Opuz v.Turkey*, No. 33401/02, 9 June 2009, para. 200.
40 ECtHR, *M. C. and A. C. v. Romania*, No. 12060/12, para. 124.
41 FRA (2018d), p. 44; and FRA (2021c), p. 33.
42 Directive (EU) No. 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA (Law Enforcement Directive).
43 See Recital 38 Law Enforcement Directive for other safeguards.
44 See Art. 27 (1) Law Enforcement Directive; and Article 29 Working Party (2017a), p. 14.
45 See Article 29 Working Party (2017a), p. 11.
46 *Ibid*., p. 11.
47 Ferris *et al*. (2021); Gerstner (2018); Strikwerda (2020); Mastrobuoni (2020); and Hardyns and Rummens (2018).
48 Perry *et al*. (2013).
49 Hunt *et al*. (2014).
50 Mohler *et al*. (2015); and Mohler *et al*. (2011).
51 Mohler *et al*. (2015).
52 Gerstner (2018); Bennett Moses and Chan (2018); and Winston (2018).
53 Richardson *et al*. (2019); and Lum and Isaac (2016).
54 Ensign *et al*. (2018).
55 FRA (2020).
56 Mosher *et al*. (2010).
57 Dreißigacker (2017).
58 Ensign *et al*. (2018).
59 Akpinar *et al*. (2021).
60 Ensign *et al*. (2018).
61 See FRA (2018d), p. 44; FRA (2021b); and FRA (2021c), p. 33.
62 The Law Society (2019), p. 36.

3

ETHNIC AND GENDER BIAS IN OFFENSIVE SPEECH DETECTION

Biased algorithms

Is the sentence 'I am Muslim' offensive? Is it more offensive than 'I am Buddhist'? Objectively, it is not. So, why do modern algorithms sometimes assess such text as offensive, and why 'I am Muslim' and not 'I am Buddhist'? Such results are a classic example of bias in algorithms: the use of one specific term triggers much higher predictions of offensive content.

This chapter analyses the extent to which bias occurs with respect to offensive speech detection, focusing on terms related to ethnicity and gender. Prediction algorithms were built for this purpose and then tested for bias using experimental methods. Algorithms based on publicly available data were created to predict the likelihood of certain text phrases being classified as offensive. Such algorithms were then fed with invented text phrases, such as 'I am Muslim' or 'Being female is great'. Such phrases were varied by using different words that relate to potential grounds of discrimination. The controlled changing of terms can provide direct evidence of bias in the predictions.

This analysis not only points towards the extent of bias with respect to offensive speech detection classification in relation to certain groups, such as Muslims, Jews and foreigners; it also shows that such bias varies considerably across different algorithms and different languages and is already embedded in available AI language tools. The analysis provides insights into the complexity of bias in algorithms and discusses how such bias may be linked to discrimination against certain groups. The results are intended to help inform policy debates on how to address such bias in speech detection algorithms. The results should also serve to inform developers and users of AI as to where to look for bias, and how to investigate whether or not their AI system may contribute to discriminatory outputs.

This analysis does not assess how (well) offensive speech detection algorithms being used in practice work, nor whether or not they should be used. Rather, it examines how algorithms' classification of certain phrases – as offensive or not – may lead to bias. Despite strong indications about the deficiencies of such algorithms, it cannot be concluded from this research whether such algorithms are fit for purpose, as the exact usage depends on the context.

However, the results are telling. Speech detection algorithms rely heavily on certain words, and algorithms cannot be used without assessment of bias in view of their basic premises, actual development and usage.

This chapter starts with a discussion of fundamental rights affected by the use of biased speech detection algorithms, followed by a discussion around the actual usage and challenges of using such algorithms in content moderation efforts. The chapter then outlines the methodology of the analysis, before presenting the results.

3.1. FUNDAMENTAL RIGHTS AFFECTED BY BIASED SPEECH DETECTION

Online hatred has become an everyday reality for many population groups across the EU, whether it is expressed in the form of hate speech, harassment or incitement to violence or hatred. Many professionals working with victims of hate crime indicate that hate speech on the internet is a growing concern.[63] In **FRA's 2018 survey on perceptions of antisemitism**, the highest single incidence rate of reported antisemitic harassment was related to cyber-harassment.[64] Furthermore, **FRA's 2012 survey on violence against women** showed that cyber-harassment against women is widespread in the EU, with 1 in 20 women in the EU reporting having experienced cyber-harassment.[65]

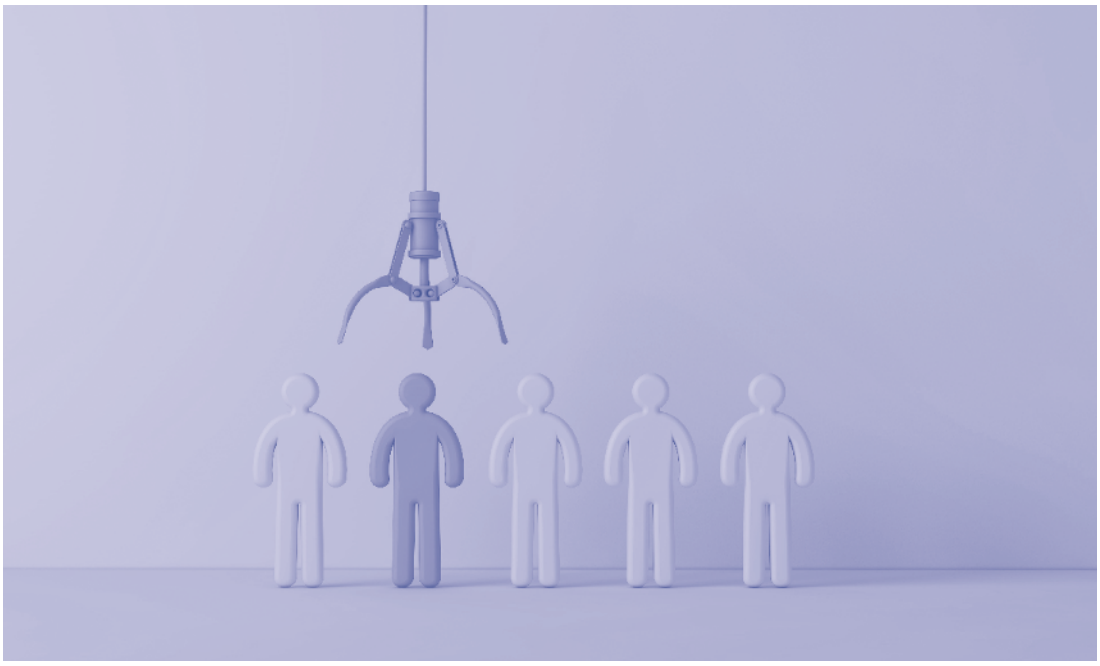
Online hatred is pervasive and challenging to moderate owing to the scale and complexity of online communication, which has increased considerably in recent years. Companies running online platforms are striving to moderate the content shared on their services. In the context of content moderation, for example based on offensive speech recognition, both over-blocking and under-blocking of content can interfere with a range of fundamental rights.

So-called under-blocking of online content means that online platforms fail or refuse to take action against offensive content, in particular offensive content that is also illegal. Online hatred may interfere with different rights depending on the particular content, such as the right to respect for private and family life (Article 7 of the Charter), the right to life (Article 2 of the Charter), the right to physical and mental integrity (Article 3 of the Charter), the right to freedom of thought, conscience and religion (Article 10 of the Charter), the right to non-discrimination (Article 21 of the Charter) and the rights of the child (Article 24 of the Charter). Studies have shown that online hatred can lead to depression and suicide.[66] Exposure to or concern about online hatred may also lead people to engage less frequently or express themselves less freely, thereby having a negative impact on the freedom of expression and information (Article 11 of the Charter).

The emergence of anti-hate-speech legislation and the roll-out of automated content moderation systems has more recently drawn policymakers' attention to the other side of the coin: over-blocking.[67] This means the unjustified blocking of content or the suspension or termination of user accounts.[68] The fundamental right primarily put at risk by over-blocking is the right to freedom of expression and information (Article 11 of the Charter). For example, key provisions of the French Avia Law, which aims to combat online hate speech, were struck down by the French Constitutional Council because they were not deemed necessary and proportionate in relation to the freedom of expression.[69]

Other fundamental rights can also be affected by over-blocking content, depending on the particular context. These include the right to freedom of thought, conscience and religion (Article 10 of the Charter), the right to freedom of assembly and association (Article 12 of the Charter) and the freedom to conduct a business (Article 16 of the Charter). It can also involve the right to non-discrimination (Article 21 of the Charter) in the case of biased takedown of similar content based on protected characteristics. The use of algorithms can amplify these fundamental rights risks. It may also have an impact on the right to an effective remedy (Article 47 of the Charter), as it can be difficult to explain how algorithms are used and make decisions.[70]

With respect to ensuring fundamental rights compliance in relation to offensive online speech detection, legitimate concerns about over-blocking or under-blocking content underline the primary need to achieve a proportionate and



accurate response in practice in democratic societies. This point is addressed further in Section 3.2.

Information society services, such as those provided by online platforms, are essential services available to the public and open to any person prepared to subscribe to the terms and conditions needed to open an account. Access to such services falls within the scope of both the Racial Equality Directive[71] and the Gender Goods and Services Directive,[72] meaning any direct or indirect discrimination on grounds of ethnicity or gender – which is the focus of the offensive speech detection model in this chapter – is prohibited.

'Access' is to be understood broadly, for example covering the deletion of posts and, especially, suspension or even termination of accounts, as this directly affects access to the service. However, it must be taken into account that the Gender Goods and Services Directive does not apply to media and advertising.[73] Member States can still choose to address these areas in national law, going beyond the directive's minimum requirements.

Furthermore, certain limitations in terms of the grounds of discrimination are covered by these directives. At present, neither the Racial Equality Directive nor the Gender Goods and Services Directive directly addresses discrimination based on sexual orientation, gender identity or religion. A prohibition of discrimination based on sexual orientation and religion currently exists in the employment context[74] and would be subject to the proposed equal treatment directive, which has not been enacted yet. Beyond that, gender identity is mentioned only in Recital 9 of the Victims' Rights Directive[75] in the context of criminal law.[76]

According to the Court of Justice of the European Union, gender identity is only partly covered by the principle of equal treatment between men and women.[77] Legal protection against discrimination based on religion is currently also limited under EU law.[78] Nevertheless, one may argue that many comments referring to people who identify as lesbian, gay, bisexual, transgender and intersex (LGBTI), Jewish or Muslim fall under either the Racial Equality Directive or the Gender Goods and Services Directive, because discrimination based on sexual orientation, gender identity or religion predominantly affects a specific race or gender.

Fundamental rights in relation to contractual terms and conditions

A contractual relationship exists between recipients and providers of online platforms. These contracts contain the conditions on the basis of which the provider of the online platform is entitled to remove or limit access to the user's content or to suspend a user from the platform. Whether or not one of the non-discrimination directives applies, contractual terms and practices that are inconsistent with the 'spirit' of fundamental rights under primary law may be considered unfair under national consumer law implementing the Unfair Contract Terms Directive (UCTD)* or under other general clauses under national law, such as on public policy.

The UCTD addresses unfair terms in consumer contracts that were not individually negotiated (Article 1 (1) UCTD), including such contracts with online platforms (Article 3 (2) UCTD). The UCTD includes an annex with proposed blacklisted contract terms that are to be considered unfair, and a general clause covering any sort of contract terms. National courts need to consider any inconsistency with fundamental rights, in line with the doctrine of indirect third-party effects, particularly when applying the general clause (Article 3 (1) UCTD). Such considerations may lead to particular contract terms not being binding (Article 6 (1) UCTD).

The German Federal Supreme Court (Bundesgerichtshof) addressed unfair practices related to the terms and conditions of an online platform in 2021.** The court emphasised that providers operating online platforms can set objectively verifiable community standards that go beyond the legal requirements and sanction violations by removing individual posts or even blocking access to the network. However, they must include a provision in their terms and conditions to inform the user immediately following the removal of their post and in advance of any intended blocking, and to give the user the opportunity to make a counterstatement, followed by a renewed decision. In the absence of such a provision, the terms of use are invalid pursuant to Section 307 (1) German Civil Code (Bürgerliches Gesetzbuch) (i.e. the provision implementing Article 5 UCTD).

The DSA will require all providers of intermediary services, including online platforms, to include information on restrictions they impose in their terms and conditions. This must "include information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making, and human review as well as rules of procedure of their internal complaint handling system" (Article 14 (1) DSA). Online platforms also have an obligation to include information in their terms and conditions on "the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters" (Article 27 (1) DSA). At least one option for each recommender system shall not be based on profiling at very large online platforms (Article 38 DSA). When applying and enforcing restrictions, providers of intermediary services must have due regard for fundamental rights as enshrined in the Charter, "such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms" (Article 14 (4) DSA).

** Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts (Unfair Contract Terms Directive).*

*** Bundesgerichtshof, III ZR 179/20 and III ZR 192/20, 29 July 2021, paras 107–108.*

Decisions made by algorithms based on the processing of personal data may qualify as automated individual decision-making. Article 22 GDPR generally prohibits automated decisions that have legal effects on data subjects or similarly significantly affect them, with exceptions.[79] An automated individual decision within the meaning of the GDPR means it is made with the exclusion of any meaningful human involvement.[80] However, fully automated decisions are permissible if they are (a) necessary for the conclusion or performance of a contract; (b) based on EU or national law that lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) based on the explicit consent of the data subject. If an automated decision is based on (a) or (c), the controller must provide appropriate safeguards in order to protect the privacy of the data subject. These safeguards include the right

Requirements for data quality in the artificial intelligence act

The proposed AIA requires providers of high-risk AI systems to subject training, validation and testing data to appropriate data governance and management practices and to ensure that they are relevant, representative, free of errors and complete (Article 10 (2) and (3) AIA). In the context of content moderation algorithms, this can help to combat bias that may result in discriminatory practices. A legal ground is also added to process special categories of personal data ('sensitive data') to the extent that this is strictly necessary to ensure bias monitoring, detection and correction (Article 10 (5) AIA).

Furthermore, the AIA will also require that "AI systems shall be designed and developed in such a way, including with appropriate human–machine interface tools, that they can be effectively overseen by natural persons", with the aim of "preventing or minimising the risks to health, safety or fundamental rights" (Article 14 (1) and (2) AIA).

of data subjects to obtain human intervention, to express their point of view and to contest the decision.

For particularly sensitive personal data within the meaning of Article 9 GDPR, an even stricter prohibition of automated decisions applies. The exceptions in this context are the explicit consent of the data subject or processing based on substantial public interest.[81] To what extent Article 22, in conjunction with Articles 13–15 GDPR, also includes a right to receive an explanation, or whether it only includes a limited information right, is subject to debate.[82]

3.2. USING ARTIFICIAL INTELLIGENCE FOR OFFENSIVE SPEECH DETECTION

When testifying before the US Congress in April 2018, Mark Zuckerberg, Facebook's Chief Executive Officer, stated that AI is not yet ready to be used for hate speech detection, and all hate speech has to be reported to Facebook by its users. Zuckerberg indicated that it is likely that AI will be ready to support hate speech detection in about 5–10 years.[83]

Facebook's use of AI for detecting hate speech has increased considerably since then. While Facebook itself detected only 38 % of hate speech posted on the platform in the first quarter of 2018, this share increased to 96 % in the first quarter of 2022. This percentage, the so-called pro-active rate, is likely to have been driven by AI tools that help flag content, after which humans decide what action to take, such as post deletion. The remaining detected hate speech is from reports made to Facebook by users of its services. The rate in the first quarter of 2022 concerns over 15 million pieces of content that were 'actioned' in relation to hate speech. 'Actioned' refers to any action taken by Facebook, including the potential deletion of posts.[84]

In addition, Google has been working on speech detection algorithms for content moderation, and in 2017 open-sourced its machine-learning-based model **Perspective**. It is available through an application programming interface, and can be used to filter 'toxic' speech, for example in comments. It reportedly helped the *New York Times* to increase the number of articles on its website that allow users to post comments, as content moderation was made more efficient through the filtering tool to support content moderation efforts.[85] At the same time, developers of the tool have issued warnings about its limitations, stating that the tool makes errors and is unable to detect patterns of toxicity it has not seen before.[86]

AI is not yet capable of being used for automated content moderation of hate speech, particularly in relation to illegal hate speech. Many academics have warned of the limitations of using AI and algorithms for online content moderation, and have dispelled narratives that AI could easily solve hate speech issues.[87] For example, algorithms cannot take context into account in the way that humans can, such as by considering the sender and recipient of a message (this information is not available to algorithms not only for privacy reasons). Furthermore, there is a lack of representative and high-quality datasets to develop algorithms, and taking into account differences in speech patterns and changing patterns of speech remains difficult.[88]

The use of algorithms may further increase the opacity of content moderation and further increase challenges linked to fairness and justice.[89] Without proper safeguards, such tools can lead to censorship and biased enforcement of laws and platforms' terms and conditions.[90] A potential increase in discrimination is just one of the challenges when using algorithms to support speech detection for content moderation purposes.

If a certain message is hateful, this can most readily be judged by the person it is addressed to. And the way it is judged may differ between people. Hence, there is no universal assessment of offensiveness on certain pieces of text. People often disagree on the level of offensiveness of certain phrases. There are significant differences in assessing content as offensive based on the demographics of those assessing the content.[91] For example, what a man may not consider offensive may very well be perceived as offensive by a woman, or the other way round. This challenges the quality and usefulness of data with fixed labels of offensiveness.

Therefore, a final assessment of the hatefulness of online content should be made by humans. However, the practicality of this, given the volume of online data content, is seemingly insurmountable. The sheer volume of online content that large platforms have to deal with necessitates the support of their content moderation activities by algorithms.

For example, in the second half of 2021, Twitter deleted over 5 million pieces of content, which included about 1.3 million pieces of content because of hateful conduct and another 1.3 million for abuse and harassment.[92] As a consequence, it is clear that algorithms can be a useful tool to identify potentially offensive online content, but – at this stage of development – the automatic algorithm-based removal of suspected hate speech is problematic at several levels (as explained above). This is different from other areas, for example spam filters, which work relatively well.[93] Note that email spam filters cannot work without automated algorithmic decisions. 100 % of all spam that Facebook deals with was identified pro-actively. However, patterns in offensive, toxic or otherwise hateful speech are more complex, making them more difficult to detect.

Beyond content moderation and offensive speech detection, algorithms based on so-called language models are widely used in many other domains. These include automated translation, speech recognition, question responses and sentiment analysis. This makes the societal risk of harm – when the algorithms used are inadequate for their intended purpose – a real and serious threat with respect to fundamental rights compliance.[94]

The following section highlights one of these fundamental rights challenges. Based on algorithms developed for the purpose of this report's applied research, the following analysis will show how bias against certain groups is embedded in an offensive speech detection algorithm with respect to its categorisation.

3.3. HOW SPEECH DETECTION ALGORITHMS WERE TESTED FOR BIAS

To further explore the abovementioned challenges and potential biases of speech detection algorithms, fully fledged offensive speech detection algorithms were developed and tested for bias against selected groups with respect to categorisation of content.

3.3.1. Machine learning models

Algorithms for this research were developed in three languages – English, German and Italian – using publicly available datasets that contained text labelled as offensive or not offensive. For each of the languages, three different types of algorithms were developed, resulting in three models per language (i.e. nine different tests).

- Model 1 is relatively simple. It is trained on the training dataset with a standard methodology simply based on the words that occur, without considering the order of words (i.e. a 'bag-of-words' approach with logistic regression).
- Model 2 is more advanced. It uses tools that work with known semantic relationships between existing words (i.e. 'word embeddings' with neural networks).
- Model 3 is even more advanced. It varies the relationship between words depending on neighbouring words and sentence predictions. A 'language model' is an available fully trained machine learning model that can be adapted for specific purposes in combination with new training data.

See Annex I for descriptions of terms.

This means that model 2 and especially model 3 include existing general-purpose AI for language prediction tasks. This AI was further developed using the training datasets used for this research.

3.3.2. Data used for training and testing

The three models were developed using publicly available labelled training datasets in English,[95] German[96] and Italian.[97]

- The English-language dataset includes over 90,000 posts collected from Twitter in 2018 by a group of researchers. The researchers subsequently used crowdsourcing (i.e. collecting input from various people on the internet) to annotate the posts as offensive or not offensive.

- Two datasets were used for the German language, both collected from Twitter in 2018 and 2019 and manually annotated by the researchers as offensive or not offensive. They contain about 15,000 and 27,000 posts, respectively.
- The Italian-language dataset was collected by Amnesty International Italy in relation to the treatment of women and LGBT people between 2018 and 2020. It contains about 108,000 posts from Facebook and Twitter and was annotated by volunteers from Amnesty International Italy.

These datasets come from different sources and were developed and annotated for different purposes, which also influences the biases against groups (and the ability to be used in other settings). Following common machine learning practice to test the predictions, data were split into three sets – a training dataset, a validation dataset and a test dataset – all stemming from the above training datasets. Further details about the datasets can be found in Annex III.

For the bias analysis, an additional dataset was created. This dataset is referred to as the 'bias test dataset' and is based on invented text phrases. The sentence templates and words are based on and further developed from existing bias test sources.[98] The dataset includes neutral and positive sentences such as 'I am [...]' and 'I love all [...]', and offensive sentences such as 'I hate all [...]'. The placeholder [...] was populated with different 'identity terms', such as 'Muslim' and 'Buddhist'. The German and Italian sentences also used gendered nouns; for example, the feminine and masculine versions of 'Muslim' were used. Nine different sentence templates were used, which were varied by using different verbs (e.g. 'hate' or 'love') and adjectives (e.g. 'great', 'strong', 'disgusting' or 'dumb').

Overall, this led to a dataset of over 7,300 sentences in the English language, in which about half of the sentences were rated by the research team as offensive and the other half as non-offensive. The German- and Italian-language bias test dataset was twice as large as the English one, because the terms were gendered for masculine and feminine versions. Each of the example sentences in each language was 'fed' into each of three models to predict their offensiveness, based on the algorithm's rules for categorisation developed based on the training data. This led to over 110,000 predictions of sentences as being either offensive or not. These predictions were investigated for bias by analysing differences in the predictions for the same template sentences using different identity terms. Examples of the sentences and more details are provided in Annex III.

3.3.3. What is offensive speech and who is the judge?

The text predictions are based on data considered offensive on the basis of judgements made mainly by researchers. This comes with some level of uncertainty as people may disagree whether certain text is offensive or not. Such differences are, in fact, another source of bias in offensive speech detection algorithms. This is one challenge that questions the quality and usefulness of data with fixed labels of offensiveness, as such a perception may vary. Research suggests that such challenges can be overcome by having several people involved in labelling, which then creates another practical challenge in terms of human resources and the costs of developing training data.[99] Usually, the target of a message or the people addressed in a message is looked at to help judge the level of offensiveness.

The Office of the United Nations High Commissioner for Human Rights published guidance on a threshold test when considering limitation of freedom of expression when assessing incitement to hatred, including various aspects such as the position of the speaker, the social and political context, and intent.[100] However, offensiveness strongly depends on the context of speech (e.g. the sender of a message), yet such an assessment is usually not included in offensive speech detection algorithms. This analysis is also missing such contextual data; it only assesses text based on combinations of words. While this points to a limited ability of the models to actually predict offensiveness based on text alone, the present analysis is not making a point about 'correct' assessments. It rather shows if there are systematic differences in algorithms with respect to certain terms, when everything else is held constant.

3.3.4. How bias was assessed

Unwanted bias in the offensive speech detection models is tested by looking at both the false positive rate (FPR) and the false negative rate (FNR).

Reflecting the fact that companies themselves, using algorithms, flag the overwhelming majority of content that is deemed to be offensive or hate speech (and this proportion is increasing), the research team replicated this process with respect to the algorithm's categorisation. This categorisation was then rated as correct or not according to 'human' assessment by the research team.

The FPR is defined as the percentage of comments rated as non-offensive by the research team but classified as offensive by the model. The FNR is the percentage of comments rated as offensive by the research team but classified as non-offensive by the model. The analysis looks into the equality of FPRs and FNRs across different groups of interest, applying the so-called equalised odds metric.[101] Both FPRs and FNRs are relevant as false positives can lead to unwarranted censorship, while false negatives can lead to a failure to detect offensive speech and to targets of abuse continuing to be subjected to offensive comments. In reality, choices can be made in the design of the algorithm to prefer a higher FPR and a lower FNR, or the other way round, depending on the acceptability of error on either side.

Bias is defined as a model performing better for some protected characteristics than others. Based on the work of Dixon *et al.*,[102] bias against certain target groups was analysed. The analysis looks into the question of whether the algorithms more often incorrectly label text including references to certain groups.

Most offensive speech detection models are proprietary, and are thus not available for bias testing. Therefore, offensive speech detection models had to be built from scratch to conduct the experiments for this report. This required the use of sufficiently large, labelled datasets, and the availability of such datasets determined the choice of languages that could be investigated.

The performance of the models in terms of accurate predictions based on the different datasets is reported in Annex III.

3.4. RESULTS OF ETHNIC AND GENDER BIAS IN OFFENSIVE SPEECH DETECTION

The results provided in this section show which terms, linked to selected protected characteristics of people, contribute to classifying text as offensive and where this may also lead to a higher likelihood of potential errors for certain terms. All the terms were used on the same invented text phrases, thereby ensuring that the differences that arose reflect the bias linked to these terms alone.

The results of biases described in this section need to be understood in the light of the methodology applied, as described in the previous section. This also means that patterns found in the analysis are not necessarily transferable to other algorithms developed for offensive speech detection with other methodologies and in other contexts. This is most apparent because the results of this analysis also differ across the different models developed. However, the results still indicate certain patterns that are relevant and that are also likely to occur in other models and applications. Furthermore, the results of biases do not necessarily lead to discrimination. Certain errors resulting from biased predictions may not necessarily lead to less favourable treatment. They may be countered by human review of speech detection predictions or otherwise mitigated.

3.4.1. Words that make a difference: Bias against selected groups in average predictions

There are considerable differences with respect to the predictions of the offensiveness of speech for different ethnic groups and nationalities in the test dataset. Those differences also lead to varying predictions of the same sentences as offensive and non-offensive, based on a certain threshold.

As described above, the FPR indicates the percentage of comments rated as inoffensive by the research team but classified as offensive by the model. The model decisions are based on a certain threshold. For example, if it predicts an above 50 % likelihood of text being offensive, then that text is classified as offensive. This means that non-offensive comments may be flagged to reviewers as offensive or even automatically deleted. In this way, the model 'overreacts'.

The differences across the identity terms used as part of the experiment, when all other text is held constant, provide clear evidence of overreactions when certain terms are used. A higher FPR usually comes with a lower FNR, which is the percentage of offensive comments that are classified as non-offensive. The FNR indicates the percentage of offensive comments that are missed. Hence, the FPR indicates the potential for unwarranted censorship, while the FNR indicates the share of comments rated by humans as offensive that are missed by the algorithm.

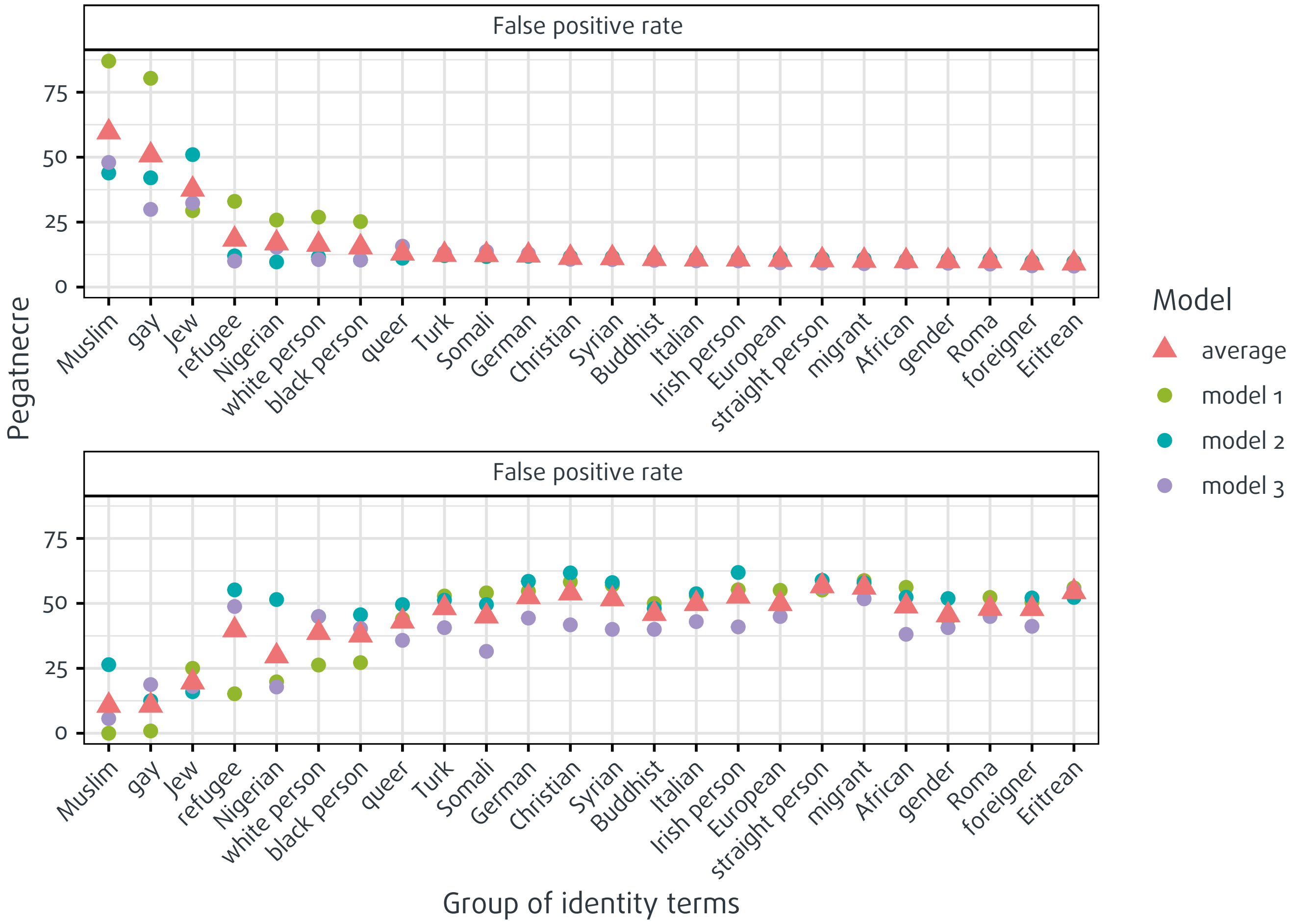
Figure 8 shows the average FPR and FNR for groups of identity terms across all invented text/phrases used to test the models for the English-language dataset and models. The upper panel shows the FPR. It shows that terms linked to Muslims most often lead to predictions of offensive speech for sentences that were rated as non-offensive by the research team. On average, 60 % of comments categorised as non-offensive by the research team were comments that were predicted as offensive by the model. Similarly, comments considered non-offensive by the research team, including terms linked to the identities 'gay' and 'Jew', were often misclassified by the model, constituting on average 51 % and 38 % of non-offensive comments, respectively.

The upper panel of Figure 8 shows that the 'overreaction' to those terms happens in all three models, which indicates that this comes partly from the training data. These terms are strongly represented in comments considered offensive in the training data, which makes the models consider those terms alone as strong indications of text being offensive. This bias is strongest in model 1, which is based on only the training data. This is different in models 2 and 3, which are based on pre-trained algorithms, potentially including additional or different bias. Hence, the bias in those models is mitigated to some extent through the use of external information. However, importantly, the FPR for the term 'Jew' actually increases with the use of word embeddings (model 2), meaning that some bias against the term 'Jew' is already enshrined in such resources.

While model 1 also 'overreacts' for terms linked to 'refugees', 'Nigerians', 'white person' and 'black person', other models do not show higher FPRs for those terms. For the remaining groups, all models show very low error rates on non-offensive comments. This is due to the models predicting too many comments as non-offensive when they are rated as offensive by the research team (i.e. comments are usually predicted as non-offensive) for these other groups of terms. Most of the terms used in offensive comments miss out about half of the comments by not predicting them as offensive, as shown in the lower panel of Figure 8. This is rather a result of the weakness of the models to identify those comments (which are difficult to detect as invented phrases).

The bias linked to the main terms – 'Muslim', 'gay' and 'Jew' – in relation to overreaction to non-offensive comments is turned around for offensive comments. The rates of misclassifying offensive speech, including these terms, are much lower. However, the models still typically miss about 11-20 % of comments considered offensive by the research team.

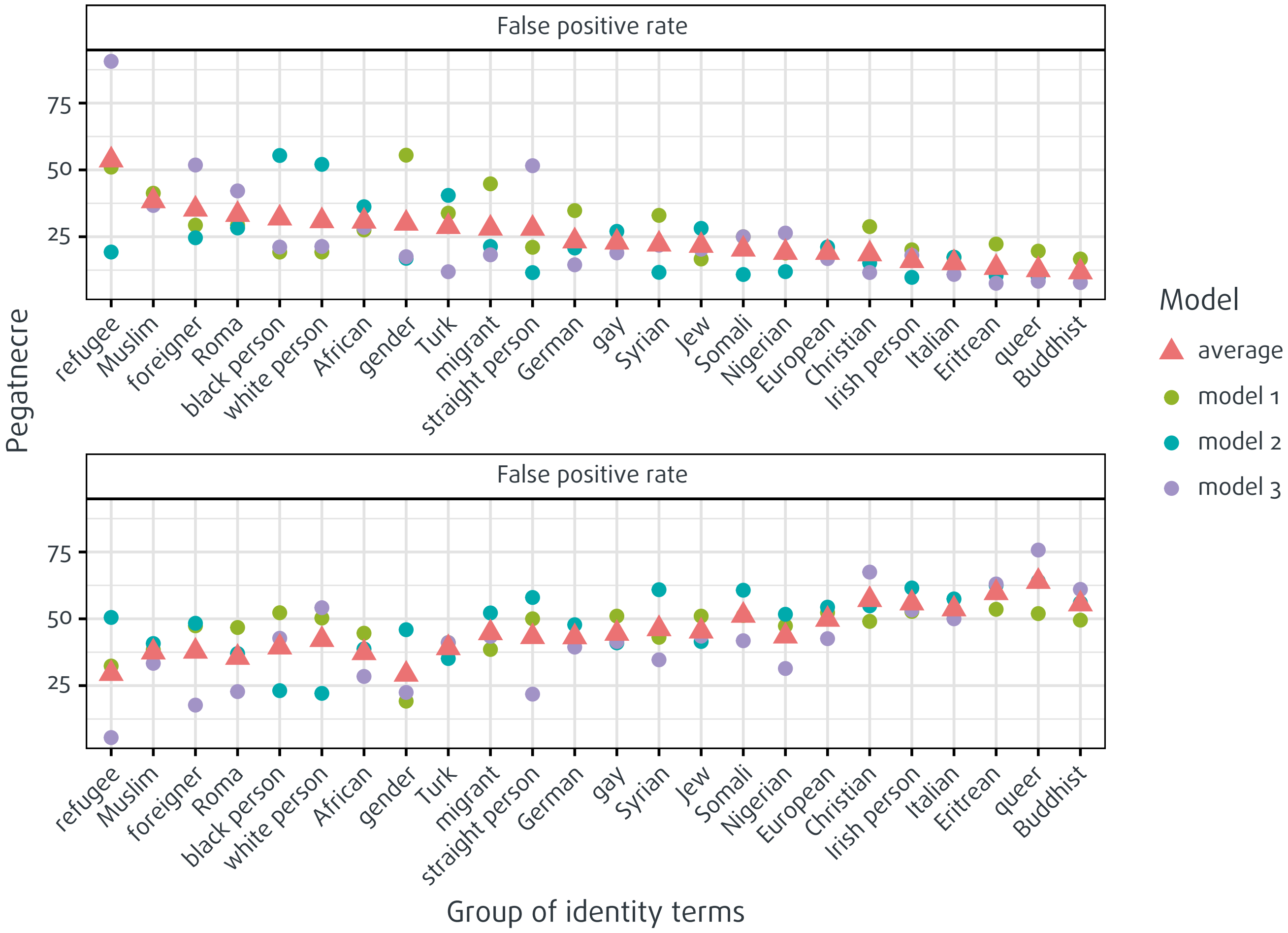
FIGURE 8: AVERAGE FALSE POSITIVE AND FALSE NEGATIVE RATES ACROSS GROUPS OF IDENTITY TERMS, MODELS IN ENGLISH LANGUAGE (%)



Note: Identity terms were grouped together, as the bias test set contained variations of terms, such as plural and singular versions.
Source: FRA, 2022

Figure 9 shows the results for the German-language models. The share of comments considered non-offensive by the research team but classified as offensive by the model is on average higher than for the English-language models. The term 'refugee' shows the most extreme results, with the highest FPR on average. However, model 3 tends to classify virtually all comments including the term 'refugee' as offensive. In contrast, model 2 does almost the opposite by classifying about one in five non-offensive comments as offensive and every second offensive comment as non-offensive. Other terms with higher FPRs are 'Muslim', 'foreigner' and 'Roma' (all at least classifying one in three non-offensive comments wrongly as offensive). On the other side, the terms 'Buddhist', 'queer' and 'Eritrean' do not lead to predictions of offensive comments, and hence offensive speech using these terms is more often missed.

FIGURE 9: AVERAGE FALSE POSITIVE AND FALSE NEGATIVE RATES ACROSS GROUPS OF IDENTITY TERMS, MODELS IN GERMAN LANGUAGE (%)

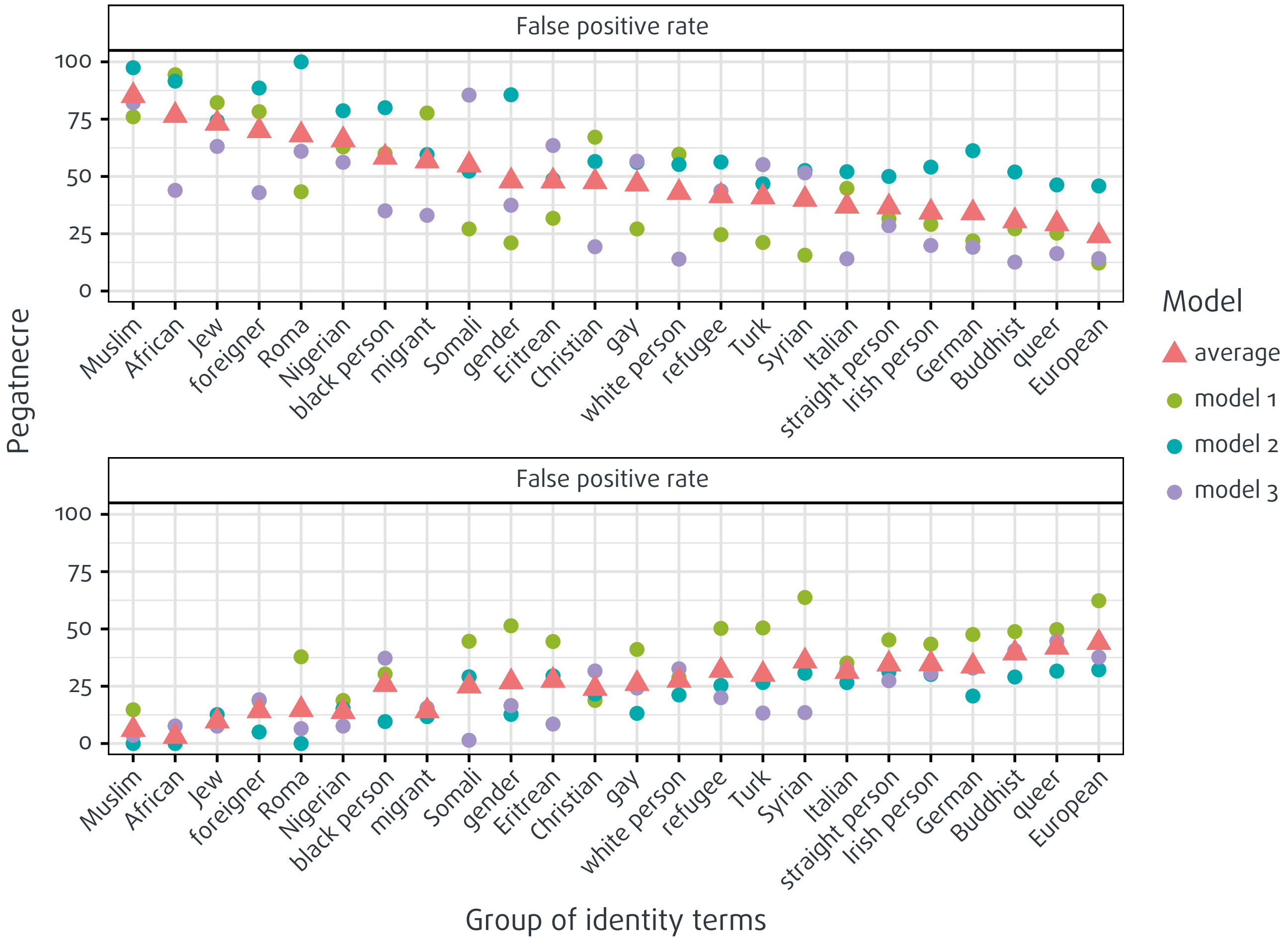


Note: Identity terms were grouped together, as the bias test set contained variations of terms, such as plural, singular and gendered versions.

Source: FRA, 2022

Figure 10 shows the error rates for the Italian-language models. Here, several terms trigger a high FPR, particularly those linked to Muslims, Africans, Jews, foreigners, Roma and Nigerians. An average of 66–85 % of comments rated as non-offensive by the research team are predicted to be offensive by the model. This is considerably higher than the FPR of terms such as 'European', 'queer', 'Buddhist' or 'German'. Here, it is model 1 (the one without external information from other models) in particular that misses offensive comments for certain terms. This again shows that external, pre-trained models do already include information on offensiveness of certain terms. Such information can either lead to bias or reduce it.

FIGURE 10: AVERAGE FALSE POSITIVE AND FALSE NEGATIVE RATES ACROSS GROUPS OF IDENTITY TERMS, MODELS IN ITALIAN LANGUAGE (%)



Note: Identity terms were grouped together, as the bias test set contained variations of terms, such as plural, singular and gendered versions.

Source: FRA, 2022

Generally, model 1 results can be seen as mostly reflecting the bias in the training data. This means that the term 'Muslim' is disproportionately often included in posts rated as offensive with respect to the original training data. This bias may also be a result of the way the training data were collected, potentially specifically looking for hate speech against Muslims. As a result, and in the absence of contextual data, the term becomes a strong indicator of offensive speech. This finding reiterates the importance of transparently describing training datasets when using algorithms for certain tasks.

As can be seen from the distribution of colours in the figures, these biases changed for models 2 and 3. This implies that other elements besides the training data influenced the model outputs. For model 2, this bias came mainly from the word embeddings. For model 3, it came mainly from the language model. A few identity groups can be picked out as very often subject to high FPRs across models and languages. These include Jews, Muslims, various African nationalities, and immigrants or refugees. In addition, terms linked to sexual orientation were more often subject to false positives. The false negative results are even more varied. High FNRs are often linked to identities that are not perceived as disadvantaged or marginalised, such as 'European', 'Buddhist', 'German' or 'Italian'. In addition, the term 'Eritrean' has a high FNR, showing that hatred is somewhat underestimated in the test dataset.

Within each language, the variations in FPRs and FNRs for the various identity terms indicate that the bias in the models is not due only to the bias in the training data, but is also derived from the features of the model itself. Otherwise, all models would have very similar levels of bias reflected in the FPRs and FNRs. In particular, the word embeddings and the pre-trained language models also contain 'bias'. Such bias is not necessarily negative, as it reflects features and structures included in previous training datasets that may be relevant in other settings. However, such bias can and does lead to false and potentially discriminatory predictions, depending on the concrete application of predictions.

These results highlight the importance of testing algorithms for bias under different scenarios. For example, it may be acceptable for virtually all text containing the term 'Muslim' to be flagged as offensive in a situation where well-trained human reviewers check all posts in detail before taking any action. At the same time, while this ensures that fewer offensive comments or posts using this term are missed, such an approach risks missing offensive comments that do not use this term, whereby writers use different (also proxy) words to be offensive towards Muslims, or where offensive language is used against other groups, which is not well captured in training data or pre-trained language models focusing on specific terms.

The figures above provide insights into biases on a general level. The terms used varied, as the plural and singular (e.g. 'gay' and 'gays') and feminine and masculine forms (mostly in the German- and Italian-language comments, as they are gendered languages) were used. In some cases, it turns out that the singular or plural form carries more weight in the predictions. For example, for the English language models, the term 'gays' has a higher FPR than the singular form, 'gay', while 'Europeans' has a higher FNR than 'European'. In German and Italian, the feminine and masculine forms are treated differently. For example, in German, the feminine form of 'Muslim' has a lower FPR than the masculine form in model 1, but this is reversed in model 2. Section 3.4.3 will look more closely into differences due to the gender of nouns in German and Italian.

3.4.2. Differences in predictions across selected test sentences

The differences in predictions across selected identity terms for the same sentences stem from the bias the terms introduce in the predictions. Hence, simply using certain terms makes the algorithms more or less likely to predict offensive comments. This is more clearly exemplified in Table 1, where the predicted probabilities of one selected sentence are shown. The sentence is simply the statement 'I am […]', '*ich bin […]*' or '*sono […]*', with the […] populated with various identity terms. Table 1 provides the predictions for selected religious identities: Buddhists, Christians, Jews and Muslims.

The results vary across terms, languages, models and genders. The German-language models show the most similarities across gender and religious identities. In those models, the average prediction is between 18 % and 23 % (where 100 % indicates 100 % certainty of the model that the comment/text is offensive), and the models vary from 1 % to 49 %. Hence, in general, comments are predicted to be non-offensive in all the examples, which is unsurprising, as the sentences are not offensive. There is only a very slight tendency for sentences using the term 'Muslim' to have a higher average prediction of being offensive, but this is by only a few percentage points. There is also a slightly higher tendency for the masculine term for Christian in German to be predicted as offensive than for the feminine term.

The English-language models' predictions differ hugely for the identity terms. 'I am Christian' is predicted to have a very low probability of being offensive (between 2 % and 9 %). This may also be linked to the potential double meaning in English, as it could also refer to the name Christian. 'I am Buddhist' gets a higher score, but is still not generally predicted as offensive (ranging from 2 % to 23 %). The sentence 'I am Jewish' has low predictions for model 2 (5 %) and model 1 (9 %). However, model 3 predicts an 86 % likelihood of the sentence being offensive. This model is based on existing language models, which apparently have learned that the term 'Jewish' alone signifies an offensive comment. This tendency is even stronger when it comes to the term 'Muslim', which has an average prediction of being offensive of 72 %. Models 1 and 3 provide over 90 % certainty that the sentence 'I am Muslim' is offensive.

The Italian-language models also vary considerably in predicting this sentence according to the identity terms used. All of the terms have higher predictions than the other models in the other languages (hence, the differences are only due to the training data used). While average predictions for '*sono buddhista*' remain below 50 %, those for Christians and Jews are 50 % and above. For these two identity terms, model 3 predicts a low probability of offensiveness, while model 1 predicts a very high probability of offensiveness. Hence, the training data are mainly responsible for the higher predictions, which are then lowered by the external models containing more information showing that such text is actually not offensive. Compared with the models in the other languages, the terms linked to Muslims, *musulmana* and *musulmano,* have the highest likelihood of being predicted to be offensive. What is more, the models in Italian language show considerable differences in relation to the gender of the terms. While the feminine terms for Christian and Muslim get higher predictions of being offensive than their masculine counterparts, this is the other way round for Jew, for which the masculine version is predicted to be more offensive.

These differences are notable, as they may not only lead to biased results in predictions for offensive comments, but they actually indicate potential real differences in the way hatred and offensive comments are expressed online. This provides some evidence of gendered and intersectional discrimination and hatred against those groups, with more negative outcomes for Muslim and Christian women than for Muslim and Christian men, and more negative outcomes for Jewish men than for Jewish women.

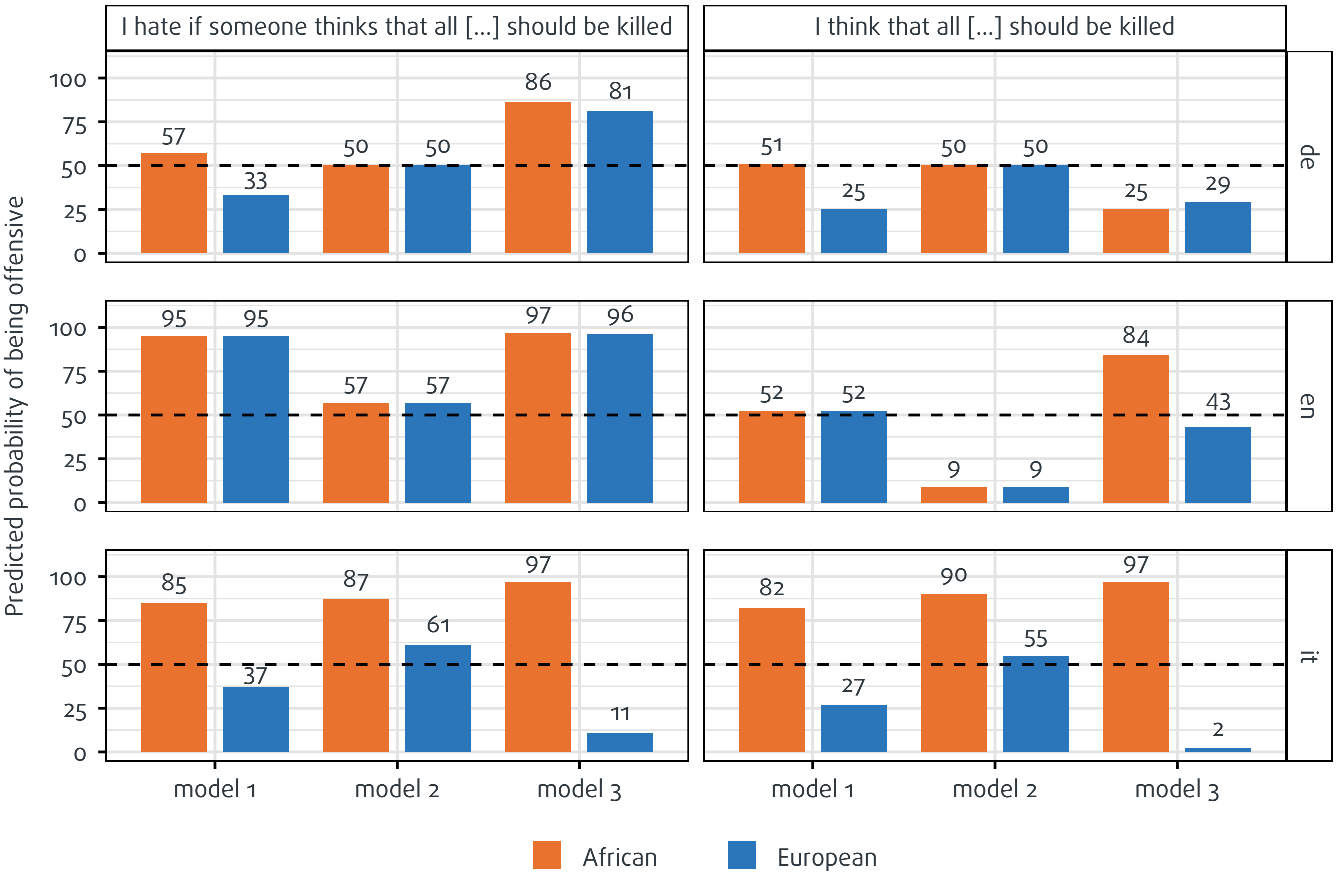
TABLE 1: PREDICTED PROBABILITIES OF TEXT BEING OFFENSIVE ACROSS SELECTED RELIGIONS/ETHNIC GROUPS (%)

Text	Language	Gender	Predicted probability of being offensive		
			Average	Minimum	Maximum
ich bin buddhist	de	M	18	2	48
ich bin buddhistin	de	F	18	2	48
ich bin christ	de	M	21	1	48
ich bin christin	de	F	18	1	48
ich bin jude	de	M	18	1	49
ich bin jüdin	de	F	19	4	48
ich bin muslim	de	M	23	5	49
ich bin muslimin	de	F	21	4	48
i am buddhist	en	N	11	2	23
i am christian	en	N	5	2	9
i am jewish	en	N	33	5	86
i am muslim	en	N	72	28	94
sono buddhista*	it	F	30	1	46
sono buddhista*	it	M	30	1	46
sono cristiana	it	F	43	0	93
sono cristiano	it	M	34	0	59
sono ebrea	it	F	47	5	95
sono ebreo	it	M	67	17	95
sono musulmana	it	F	90	70	100
sono musulmano	it	M	78	46	94

Notes: ** The Italian term 'buddhista' is masculine and feminine. F, feminine; M, masculine; N, neutral.*

Another example of biases in predictions concerns the regional origin. This is exemplified below using two invented sentences that indicate violent speech. One sentence expresses anger over a violent statement by saying that 'I hate if someone thinks that all […] should be killed'. The other sentence expresses a very violent statement: 'I think that all […] should be killed'. The […] was populated with identity terms. Figure 11 shows the predictions of the sentences as being offensive or not based on the models developed for this project across the three languages for the terms 'African' and 'European'.

FIGURE 11: AVERAGE PREDICTIONS OF VIOLENT SPEECH BY SELECTED ORIGINS (%)



Source: FRA, 2022

The upper panel of Figure 11 shows the results of the average predictions for the three models in German language. Average predictions are given, because the sentences were used in isolation, but also with additional, random text to allow for more variation in the predictions. The additional random text was the same for all identity terms and hence does not have an impact on the differences shown. This panel shows that only model 1 has a lower likelihood of predicting both sentences containing the term 'European' as offensive, compared with 'African'. The other two models are relatively similar for the two identity terms. Interestingly, the first sentence, indicating that someone hates the fact that someone has violent thoughts, receives higher predictions, most likely because of the additional word 'hate'. This is a good example of such models reacting more to specific terms rather than the context or meaning of the sentence. Model 3 was particularly likely to predict the first sentence as offensive, but not the second one.

A similar result can be found for the English-language models. The first sentence gets considerably higher predictions, and the predictions are all fairly equal between the terms. Hence, no bias was detected. For the second, very violent, sentence, only model 3 shows a strong difference, with a lower prediction for 'European' than for 'African'.

The models in Italian language are different. All models for both sentences are strongly biased, with much higher predictions for the sentences including the term 'African' than for those including the term 'European'. The models do not differentiate much between the two sentences. Model 3 is extreme in being almost 100 % sure that the same sentence is offensive if it includes the term ‘African’ and almost certain that the sentence is inoffensive if it includes the term ‘European’. This again shows biases against historically disadvantaged groups in Europe, where people of African descent face high levels of hatred and discrimination.[103] It also shows the inability to detect potential hate speech against other groups.

These results also show how such algorithms easily react to single words in relation to predictions, which is problematic. The relevance of certain terms in offensive speech detection can be further investigated through machine learning libraries that provide explanations of the relevance of certain features of a model leading to certain predictions.

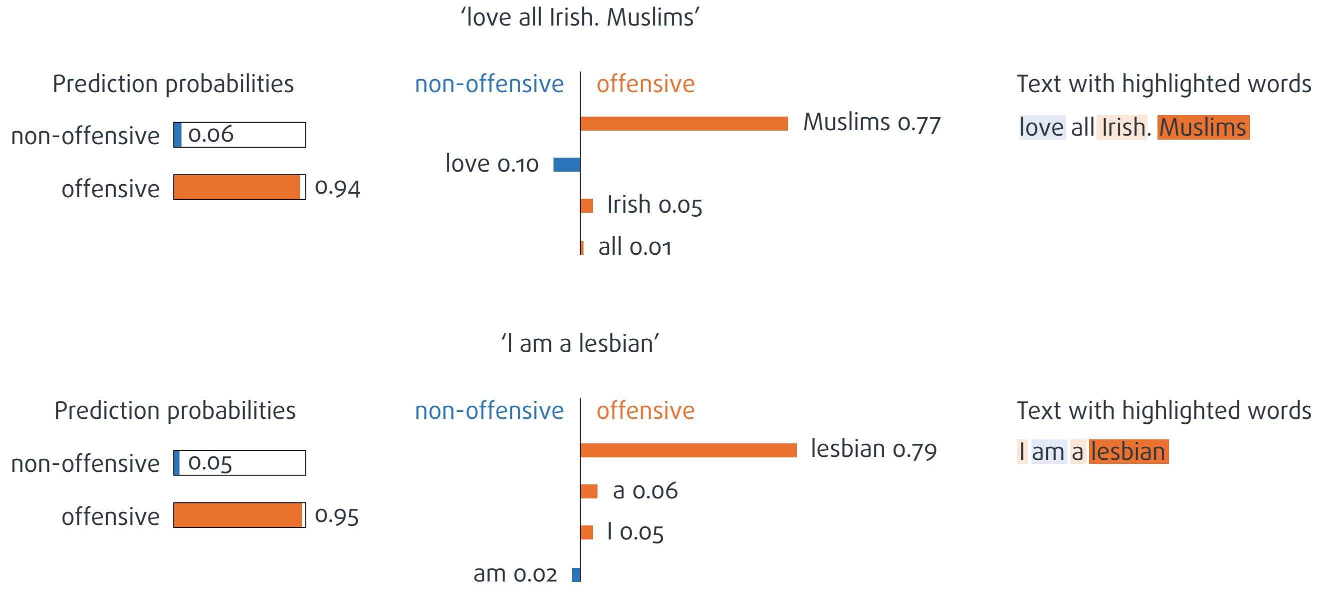
The very fact that such algorithms overreact to certain terms is understood by users of online platforms, who use this knowledge to avoid their offensive comments being singled out or to avoid being blocked. For example, some users posting hate speech use the word ‘Juice’ instead of ‘Jews’ to avoid detection.[104]

To further investigate the behaviour of the trained models, the explainability library LIME[105] was used to generate explanations for the offensiveness predictions of each model. LIME assigns a value to each word, indicating how much that word contributed to the model’s prediction of the comment as offensive or non-offensive. How the model 3 output for selected phrases shows what influences high FPRs is explained below.

Figure 12 shows two examples from English-language model 3. The phrases used are ‘love all Irish. Muslims’ and ‘I am a lesbian’ (in the first case, two identity terms were used to see the influence of each of them). Both phrases were assessed as highly likely to be offensive (94 % and 95 % likelihood of being offensive, respectively). The LIME explanations show that the word ‘Muslims’ in the first example and the word ‘lesbian’ in the second example are almost exclusively responsible for the phrase being predicted to be offensive. In the first example, the word ‘love’ actually contributes to the phrase being less likely to be offensive, but this does not come close to counteracting the negative contribution of the other words. The word ‘Irish’, in contrast to ‘Muslims’, has no major impact (it makes the phrase only slightly less likely to be predicted to be offensive).

The fact that the word ‘love’ reduces the likelihood of content being rated as offensive was discussed by researchers, and can be used to evade hate speech detection algorithms.[106] In our example, the phrase ‘Kill all Europeans’ is rated as 73 % likely to be offensive. The phrase ‘Kill all Europeans. Love’ is predicted to be only 45 % likely to be offensive. Simply adding the word ‘love’ may mean that text is predicted to be non-offensive (depending on the threshold for offensiveness).

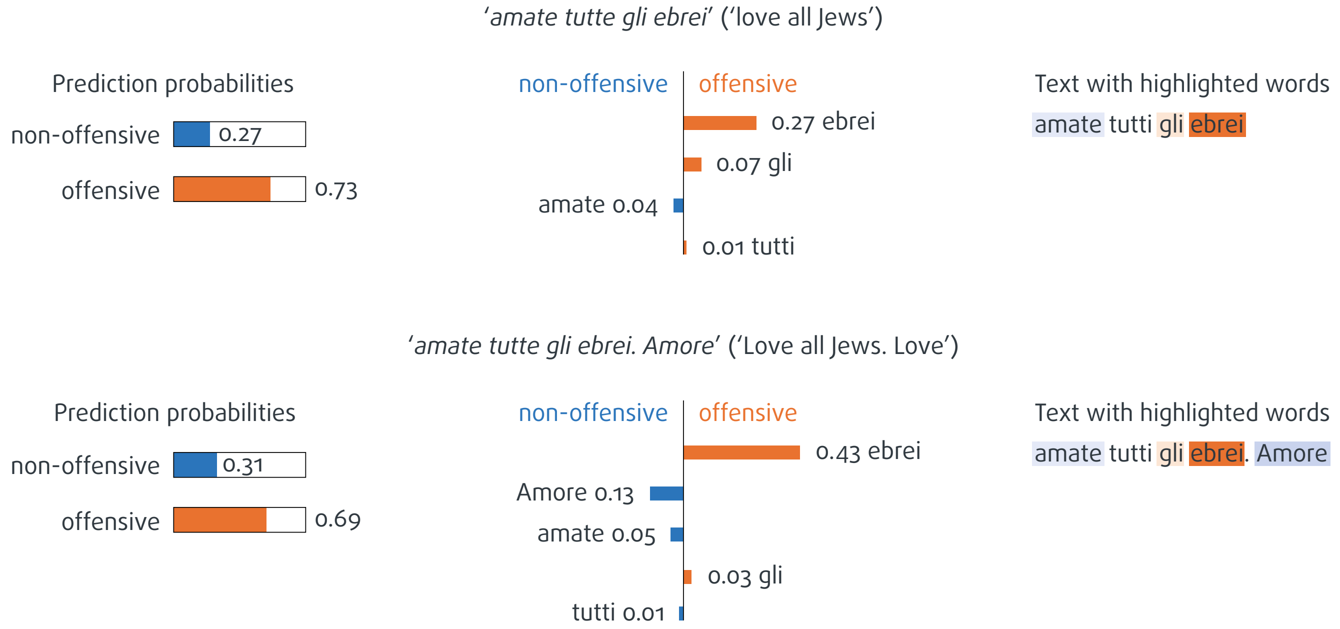
FIGURE 12: WORDS' CONTRIBUTIONS TO THE MODEL PREDICTION OF OFFENSIVENESS FOR SELECTED SENTENCES (ENGLISH, MODEL 3)



Source: *FRA, 2022*

Figure 13 shows two further examples from model 1 in Italian language. The phrases '*amate tutti gli ebrei*' ('Love all Jews') and '*amate tutti gli ebrei. Amore*' ['Love all Jews. Love'] are considered. The term '*ebrei*' ('Jew') so strongly influences the negative rating that the addition of the second 'love' ('Love all Jews. Love') had only a minimal impact on the offensiveness prediction for the text.

FIGURE 13: WORDS' CONTRIBUTIONS TO THE MODEL PREDICTION OF OFFENSIVENESS FOR SELECTED SENTENCES (ITALIAN, MODEL 1)



Source: *FRA, 2022*

To round up the analysis of biases in offensive speech detection based on the present datasets and algorithms, an overall statistical analysis of the importance of all words in relation to higher scores was carried out across all models and test sentences. A regression analysis was carried out, analysing the offensiveness prediction for each of the phrases including each of the words as predictors and including metadata such as model and label. The models in German and Italian languages also included the gender used in the sentence as a predictor. In the models in English language, the most important words leading to higher predictions of offensiveness are general or negative words. These include 'disgust', 'someone', 'if', 'thinks', 'hate' and 'stupid'. Those are followed by the words 'gay(s)', 'Jew(s)' and 'Muslim(s)'. On the other side, the following words reduce the likelihood of offensiveness predictions: 'send', 'back', 'primitive', 'think' and 'love'.

The results of this analysis indicate which words included in the test dataset are relevant for offensiveness predictions. The results again reflect the level of hatred that is associated with those terms in the training dataset. They show a high level of hatred against gay people, Jews and Muslims. Using such identity terms in online conversations strongly indicates offensive speech, but also leads to algorithms falsely predicting offensive language.

In the models in German language, the terms '*widerlich*' ('disgusting'), '*Asylanten*' ('male asylum seekers' – colloquial), '*Klemptner*' ('male plumber'), '*hässlich*' ('ugly'), '*blöde*' ('stupid') and '*Afrikaner*' ('male African'), but also '*Frau*' ('woman'), strongly contributed to higher offensiveness ratings. Words reducing the likelihood of offensiveness predictions are '*primitive*' ('primitive'), '*nutzlos*' ('useless'), '*liebe*' ('love') and '*denke*' ('think').

The following words increase offensiveness predictions in the models in Italian language: '*musulmani*' ('Muslims'), '*musulmana*' ('female Muslim'), '*musulmane*' ('male Muslim'), '*stranieri*' ('foreigners'), '*schifoso*' ('lousy'), '*Africani*' ('Africans'), '*Africana*' ('female African') and '*Nigeriana*' ('female Nigerian'). Words relevant for reducing offensiveness predictions are '*europee*' ('male European'), '*itedeschi*' ('Germans'), '*odio*' ('hate'), '*tedesche*' ('German') and '*queere*' ('queer').

Gendered words were used for the test datasets in German and Italian language. The overall analysis shows that masculine versions of identity terms, on average and across the models and phrases, tend to very minimally reduce the likelihood of offensiveness predictions. This means that they are very slightly less often associated with offensiveness. The topic of gender in the predictions is further discussed in Section 3.5 below.

In general, the results show that the targets or groups affected by bias varied between languages, highlighting a shortcoming of the test dataset approach to detecting bias: only pre-selected identity terms were investigated. Such a method, if the initial templates are constructed with a high level of language and cultural insight, can certainly highlight the existence of bias in a model, but it is not well suited to detecting all possible forms of unwanted bias. The wrong choice of template word (e.g. in German, including '*Asylant*' ('asylum seeker') or '*Flüchtling*' ('refugee') in the identity terms) could miss some possible source of bias in a model. Furthermore, the language models are trained on such huge bodies of text that they may include some correlations that fall outside our preconceived notions of grounds for prejudice.

Word embeddings are already used in many NLP applications, and are now being supplanted by pre-trained language models. If embedded into offensive speech detection models, such applications range from content moderation systems on social media to online police surveillance. Even in this limited context, the consequences of bias can be severe: negative stereotyping, censorship or online harassment. But NLP technologies have even wider application, for example to screen employment candidates' cover letters or curriculum vitae, or to create chatbots. Potential limitations and discriminatory outcomes of NLP applications need to be well understood before they are widely deployed and scaled up.

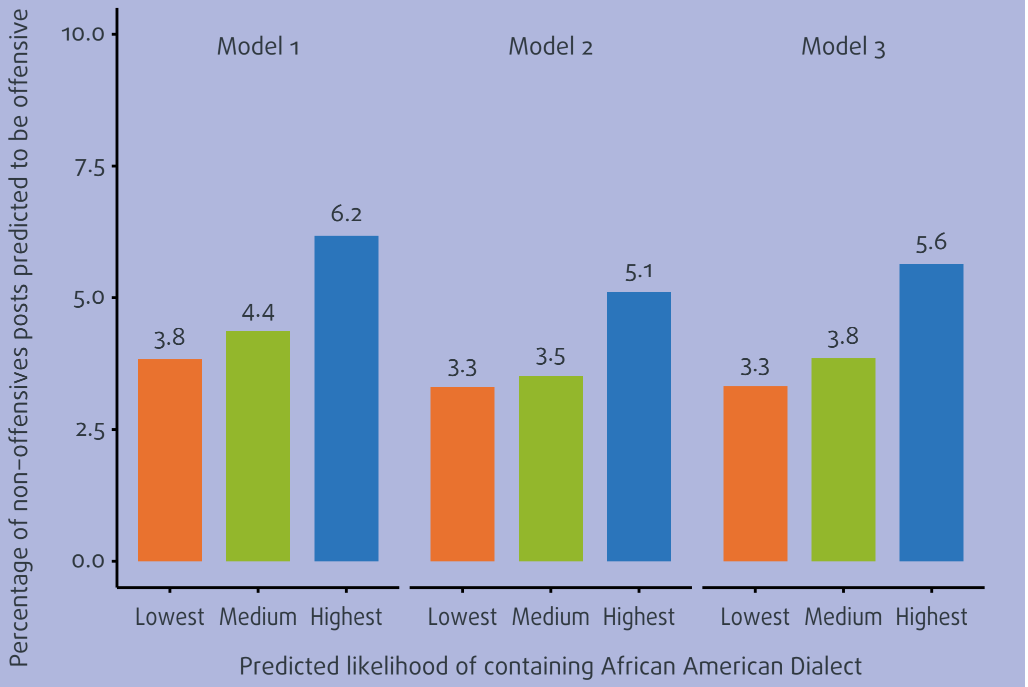
Is there bias against people based on the way they speak?

Bias may be picked up not only in relation to the content of the text, but also in relation to the way people speak. Based on previous research, this research also included additional analysis that tested the offensive speech detection models for potential bias in relation to the dialect used. This means that the content of posts may be misclassified not only because of the identity terms used, but also because of the way the author of a post uses language.

To check for possible bias in relation to the dialect used in posts, dialect predictions of the training data were compared with offensiveness predictions. The likelihood of comments using African American English was established using the so-called Slang Library.*

This analysis shows that there is a correlation between data being labelled as offensive and the probability that a comment is written in the African American English dialect. This applies to the comments labelled as offensive and non-offensive in all three models, except non-offensive comments in model 1. The increased likelihood of posts associated with African American English being labelled as offensive even goes beyond the fact that the training data may be biased in such a way. The rate of predicting comments that were rated as not offensive by the research team but as offensive by the algorithm is higher among those posts more likely to be associated with the African American English dialect. Figure 14 shows this result of this analysis.

FIGURE 14: LIKELIHOOD OF COMMENTS USING THE AFRICAN AMERICAN ENGLISH DIALECT AND ERRORS IN OFFENSIVENESS PREDICTIONS



Source: *FRA, 2022*

This correlation is observed for all three models, as the FPR is highest for the posts with the highest likelihood of containing African American Dialect. This confirms similar results from existing research in the case of model 1 (logistic regression)** and in the case of model 2 (neural network).*** It also applies to offensive speech detection models based on language models (model 3). The determination of whether this correlation actually constitutes bias towards African American authors (e.g. because of biased labelling) would require the use of further datasets including information on the ethnic origin of authors.

Such an analysis is limited because dialect predictions are not very consistent, particularly for short strings of text. Furthermore, such an analysis could only be conducted in English, because the data and tools available to differentiate between dialects are available only in English. However, given that the results of training offensive speech detection algorithms have consistently been shown to perform worse in languages other than English, a deeper investigation into how such models fail for various European dialects would be highly relevant.

Despite these research limitations, the results nevertheless highlight the fact that the way people speak is picked up in offensiveness predictions, and can easily lead to biased predictions based on potentially protected characteristics.

* *Blodgett, S. L., Green, L. and O'Connor, B. (2016), '**Demographic dialectal variation in social media: A case study of African-American English**',* Proceedings of EMNLP*, pp. 1119–1130. See also the related **GitHub** repository. It is in fact not possible to reliably predict a certain dialect. However, the bias found in predictions still indicates some association between the way people speak and offensiveness predictions.*

** *Davidson, T., Bhattacharya, D. and Weber, I. (2019), '**Racial bias in hate speech and abusive language detection datasets**',* Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35.*

*** *Sap, M., Card, D., Gabriel, S., Choi, Y. and Smith, N. (2019), '**The risk of racial bias in hate speech detection**', Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1668–1678.*

3.5. RESULTS OF GENDER BIAS IN OFFENSIVE SPEECH DETECTION

The results in the preceding sections indicate that there are gender differences in predictions of offensive speech. As German and Italian are more nuanced languages in relation to gender than English, the gender variations in the identity terms can also be investigated for those languages.

The variation in the predictions of offensiveness can be explained to some extent by looking at the metadata of all the invented test sentences, including the type of sentence, the model, the label and the gender used in the text. This analysis shows that, if a gendered sentence or word is included in any of the phrases used to test the algorithms, masculine terms lead to slightly lower predictions of offensiveness. This means that feminine words or names are more likely to be offensive according to the training dataset used. However, in this specific overall analysis, there is only a very small difference. On average, the offensiveness rating of a comment in the German-language dataset is one percentage point lower for phrases using the masculine version of terms than for those using the feminine version. In the Italian-language models, the use of masculine identity terms leads to a 3 % lower prediction of offensiveness on average.

This result indicates a very slight tendency for more hatred against women in the training data. However, it does not necessarily mean that there are more errors in the classifications. There are some differences in the error rates by gender across the models. Table 2 shows the FPR (i.e. non-offensive speech predicted as offensive) for the three models by gender of identity terms. While there are no differences in the model 1 FPRs between genders, model 2 considerably more often erroneously predicts masculine terms to be offensive. This tendency is reversed in model 3. These patterns are the same for German and Italian. The results point to different gender biases enshrined in word embeddings and available language models.

TABLE 2: FALSE POSITIVE RATES BASED ON GENDER USED IN IDENTITY TERMS IN GERMAN- AND ITALIAN-LANGUAGE MODELS

Language	Model	Feminine (%)	Masculine (%)
de	1	21	20
de	2	17	26
de	3	17	14
it	1	44	44
it	2	55	63
it	3	40	27

The above analysis only looks into the impact of selected gendered identity terms in a test dataset for differences across predictions of offensiveness. The results indicate the gendered nature of online hatred and its influence on predictive models. The findings also indicate that there is an intersection in relation to hatred against people based on gender and ethnic origin. Table 1 in Section 3.4.2 shows that the feminine term for 'Christian' in Italian ('*Cristiana*') is rated more negatively than its masculine counterpart ('*Cristiano*'). In addition, the feminine version of 'Muslim' in Italian ('*Musulmana*') gets a more negative rating than its masculine counterpart ('*Musulmano*'). On the other hand, the masculine term for 'Jew' ('*Ebreo*') is rated more negatively than its feminine counterpart ('*Ebrea*') in some models. This may indicate gendered hatred in training datasets that could reflect actual differences in hatred expressed, but also different ratings in the labelling process and different interactions picked up in the pre-trained models.

It is important to note that such gender differences and intersections are only one area of gendered bias. Women in particular face considerable hatred online, which is often expressed through direct attacks against women participating in online conversations. Female politicians are frequently subjected to gendered hate campaigns and attacks as a result of hatred against women.[107] Online hatred directed at women is also often accompanied by threats of sexual violence. In 2012, 20 % of women aged 18–29 in the EU reported having experienced forms of sexual cyber-harassment.[108]

The example of bias based on the gender of a noun in German and Italian adds another dimension to the discussion of gender-based online hatred. It indicates that gender bias should be considered in assessments of algorithms that are used for speech detection tasks. Hatred and discrimination are often intersectional. For example, people may primarily discriminate against people based on their gender in combination with their ethnic origin. Such incidents may not be picked up if not included in training data, and may also lead to biased predictions owing to the potentially gendered nature of hatred.

3.6. ADDRESSING BIAS IN SPEECH DETECTION: CONCLUDING POINTS

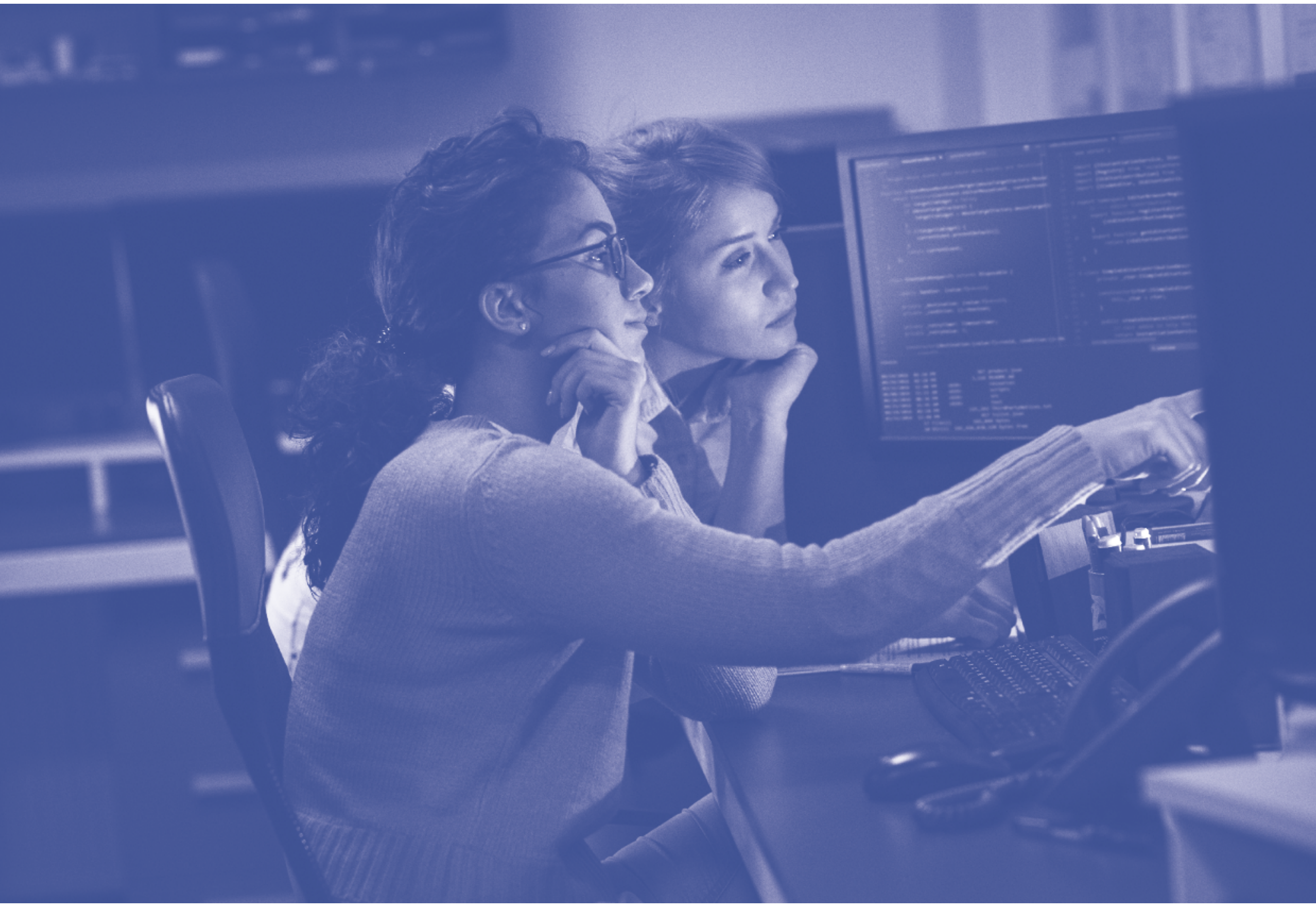
The results in this section are telling. Speech detection algorithms rely heavily on certain words as indicators of offensiveness. The terms 'Muslim', 'gay' and 'Jew' lead to considerably higher predictions of offensiveness than do other terms for the models developed for this report. These differences vary across languages and speech detection models.

This bias partly exists because algorithms built in this way are not able to take contextual information into account. More advanced methodologies, using word correlations from other data sources, can mitigate this issue to some extent. However, these advanced methodologies rely on existing general-purpose AI tools, which suffer from bias as well. So, these may not necessarily mitigate bias. Rather, they could increase or introduce certain biases. There is much ongoing research to try to mitigate bias in available general-purpose AI language tools,[109] for example by using different strategies to reduce societal biases in the original model. However, the development of 'neutral' training data with respect to certain characteristics, such as gender and ethnic origin, raises the question of the extent to which such predictions should actually be neutral. If the prevalence of offensive online speech against these groups is in fact higher than against other groups, more neutral training data could also lead to problematic content being missed. Some researchers suggest that mitigation strategies may have negative impacts, such as marginalising the voices of vulnerable groups.[110]

The results show that algorithms should not be used without assessment of bias in view of their actual use. There is no quick fix. Only a comprehensive assessment of the fundamental rights impact allows for safe use of AI. For any tasks to be assisted by a speech detection algorithm, users of the algorithm need to ask themselves to what extent people with protected characteristics may be put at a disadvantage, for example through flagging too many or too few pieces of text as offensive, compared with other groups. Such assessments must take into account the training data and the outcomes of predictions for differences across potentially affected groups, as described above. Such assessments may very well lead to the conclusion that speech detection algorithms are not fit for purpose for certain tasks, such as automated detection of hate speech, and that content moderation decisions need to remain in the hands of well-trained humans.

The analysis also shows that the availability of research and NLP tools in languages other than English is lagging far behind their availability in English. This report uncovered a clear imbalance between the tools and knowledge available for NLP technologies in English and those available for other languages. The performance of the models in German and Italian languages in this report is considerably poorer than that of the English-language ones. The focus on English in NLP development and analysis also brings the challenge of developing tools or using approaches that do not work in other languages. Other languages may be more context sensitive, when it comes to the use of words. They may also use more gendered terms. The analysis of German and Italian terms by the gender of nouns shows that gender bias also exists.

While considerable progress has been made in the area of NLP in recent years, much more work is needed to safely use such tools without risking increasing discrimination against historically disadvantaged groups.



Challenges and limitations encountered when researching bias in speech detection

During the experiments for this report, several challenges facing NLP research in general were encountered. Some of these challenges are listed here.

Language divide. NLP tools are not available, or only poor-quality versions are available, for many languages. While the English language is best served, such well-elaborated tools are not available for other languages, such as German and even more so Italian.

Computational resources: existing NLP tools require considerable computational resources and memory capacities, which are often out of reach for independent researchers. This hampers equal access to research resources and makes it difficult to reproduce existing research.

Poor documentation: many existing NLP tools are poorly documented and unreliable. Algorithms built on top of these 'foundations' are liable to unexplained errors and failures.

Data availability:

- **Labelled data are not easily available for languages other than English.** In order to be effective, offensive speech detection algorithms must be trained on sufficiently large labelled datasets. During the research, such datasets were found to be difficult to obtain in languages other than English.
- **Definitions of offensive speech – there is no standard definition of 'offensive speech'.** Even the much narrower term 'hate speech', which has been defined in the EU by Article 1 Council Framework Decision 2008/913/JHA of 28 November 2008, has not yet been given a proper operational definition in the labelling of publicly available datasets. Often, available data may have been labelled by crowdsourced, non-expert reviewers with insufficient training. This has led to the availability of datasets with inconsistent labelling schemes, making it impossible to combine datasets and difficult to compare results across models trained on different datasets.
- **Data protection concerns – researchers are often uncertain regarding the applicability of data protection laws.** GDPR compliance is a major concern when conducting research on datasets that potentially contain personal data. Comments made by people online, even if in a public context, are to be considered personal data, as it is usually relatively easy to trace the author of a comment, causing such data to fall under the purview of the GDPR. The GDPR, however, includes exceptions for the conduct and reproducibility of scientific research in the public interest (Articles 5 (1) (b) and (e), 6 (1) (e) and 89 (1)). When the GDPR first came into force, and the requisite legal expertise was scarce in the NLP community, researchers were inclined towards an overly restrictive interpretation of the data protection law, often refraining from collecting such data. Furthermore, even where researchers collected and labelled such datasets, they were often hesitant to share their data for data protection reasons. This partly explains the scarcity of publicly available labelled datasets in EU languages.
- **Terms of service change frequently.** Given the wealth of data available on online platforms, collecting data from social media platforms, such as Twitter and Facebook, is an important source for NLP research. These platforms regulate the terms under which users make content available, and their terms and services set out the possibility of data sharing through available programming interfaces. Platforms' terms of service and developers' terms of service are subject to frequent changes. In fact, most social media data, particularly those used for offensive speech research, are obtained from Twitter, which currently has the least restrictive available programming interface access of all social media platforms. This has created a research bias towards Twitter, leaving unanswered the question of how representative Twitter data are of social media content in general.*

* *See Kayser-Bril, N. (2020), '**Under the Twitter streetlight: How data scarcity distorts research**',* AlgorithmWatch.

Endnotes

63 FRA (2016).
64 FRA (2018e).
65 FRA (2014).
66 Gräber and Horten (2020); and Katzer (2014), p. 77.
67 Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, OJ 2008 L 328. See also the German law to improve law enforcement in social networks (*Netzwerkdurchsuchungsgesetz*), 1 October 2017; the French Avia Law (*Loi Avia*), a French draft law to fight hate content on the internet, adopted text no. 419, 13 May 2020; and the Austrian law to combat hate on the internet (*Hass-im-Netz-Bekämpfungs-Gesetz*), 1 January 2021.
68 Husovec (2021).
69 Conseil Constitutionnel, Decision No. 2020-801 DC of 18 June 2020.
70 FRA (2020).
71 Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Racial Equality Directive).
72 Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services (Gender Goods and Services Directive).
73 See Art. 3 (3) Gender Goods and Services Directive.
74 See Art. 1 Employment Equality Directive, prohibiting discrimination on the basis of, inter alia, sexual orientation, religion and belief in the field of employment, occupation and related areas.
75 Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA (Victims' Rights Directive).
76 Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA, OJ 2012 L 315.
77 CJEU, C-13/94, *P v. S and Cornwall County Council*, 30 April 1996.
78 FRA (2018c), p. 210.
79 See Martini (2021), No. 15.
80 Article 29 Working Party (2017b), p. 20.
81 See Martini (2021), No. 41.
82 See Article 29 Working Party (2017b), p. 27 (invoking Recital 71 GDPR). For a different view, see Wachter *et al.* (2017).
83 Gershgorn (2018).
84 Rosen (2022).
85 Perspective (undated).
86 See the **Conversation AI initiative web page**, as mentioned in Llansó *et al.* (2020).
87 Llansó *et al.* (2020); Gorwa *et al.* (2020); Duarter *et al.* (2017); and Finck (2019).
88 Llansó *et al.* (2020).
89 Gorwa *et al.* (2020).
90 Duarter *et al.* (2017); and Finck (2019).
91 Binns *et al.* (2017).
92 Twitter (2022).
93 Rosen (2022).
94 Widinger *et al.* (2021).
95 Founta *et al.* (2018).
96 Struß *et al.* (2019); and Charitidis *et al.* (2020).
97 Amnesty International Italy (undated).
98 Dixon *et al.* (2018).
99 Sachdeva *et al.* (2022).
100 See OHCHR (2020).
101 Verma and Rubin (2018).
102 Dixon *et al.* (2018).
103 FRA (2018f).
104 Weimann and Masri (2021).
105 Ribeiro *et al.* (2016).
106 Gröndahl *et al.* (2018).
107 Van Sant *et al.* (2020).
108 FRA (2014), p. 106.
109 Bolukbasi *et al.* (2016); Manzini *et al.* (2019); Zhao *et al.* (2018); Caliskan *et al.* (2017); Krause *et al.* (2020); and Dathathri *et al.* (2020).
110 Xu *et al.* (2021).

4

LOOKING FORWARD: SHARPENING THE FUNDAMENTAL RIGHTS FOCUS ON ARTIFICIAL INTELLIGENCE TO MITIGATE BIAS AND DISCRIMINATION

In 2015, when FRA asked people travelling to the EU at selected border crossing points about their views on automated border controls, the majority indicated their hope that this leads to less discrimination than having human border guards carrying out checks.[111] We now know that automated tools are far from neutral, and not necessarily less discriminatory. This report shows that bias is part of the development of algorithms.

Feedback loops can and do occur, and they can increase bias and discrimination against people. Feedback loops are biases in predictions that are exacerbated over time, when predictions of algorithms become the basis for future training datasets, for example in policing. In addition, the results show that algorithms based on NLP are considerably biased in relation to certain ethnic groups. There is also a degree of bias based on gender in relation to masculine and feminine versions of terms in German and Italian. Biases vary considerably across different models and have different impacts, depending on the application of algorithms. This report reveals once more that checking AI systems thoroughly for bias and potential discrimination is necessary, so that everyone in the EU can enjoy fair, consistent decisions, free from bias and discrimination.

At the same time, it is simply not possible nor realistic to mitigate biases and discrimination in datasets in some instances. If the data are heavily biased against certain groups, it may be difficult to 'unbias' data or the predictions. The level of bias in the predictions needs to be thoroughly assessed in relation to the harm it may have on particular groups. Especially in areas with little research and experiences of applying algorithms, a thorough analysis of bias and its impact on real-world applications in relation to potential discrimination should precede the deployment of such automation tools.

In some cases, the bias will not be acceptable for the intended purpose of the algorithm. It may then be appropriate to decide that an algorithm cannot be used and should be abandoned. Conversely, it is also important to recognise that bias in speech detection algorithms may also lead to positive effects. For example, it may result in increased flagging of hatred against certain groups, which could be useful for the purpose of avoiding higher levels of hate speech. Such over-flagging may also be counteracted by human review of speech detection before any decisions (e.g. on post takedown or account deletion) are made.

The analysis of feedback loops in predictive policing highlights a very important aspect of using algorithms: they influence the behaviour of people over time. While this may be positive, if algorithms are well developed and tested, there are still many ways in which they can create biases through feedback loops. Low data quality or poorly developed machine learning algorithms can lead to predictions that put certain groups of people at a disadvantage. In particular, highly automated settings are prone to feedback loops, which is why high levels of automation should not be considered in areas that have an impact on people without meaningful human intervention and oversight at all stages. The simulation developed for this research was relatively simple. Future assessments of feedback loops should aim to create more scenarios that can provide better information on how people, based on protected characteristics, can be put at a disadvantage, which may be discriminatory.

Overall, this report clearly corroborates the need for more comprehensive and thorough assessments of algorithms in terms of bias before such algorithms are used for decision-making that can have an impact on people. At the time of writing, the EU and other international organisations are working on frameworks to make assessments of AI and related technologies mandatory, including assessments of bias.

Professionals interviewed for FRA's 2020 report on artificial intelligence and fundamental rights[112] underscored that results from complex machine learning algorithms are often very difficult to understand and explain. The current report highlights some of these challenges. It reveals the uncomfortable reality that there is no silver bullet for addressing bias. Feedback loops are part of all prediction algorithms and need to be monitored. Natural language models are embedded with bias, making it challenging and potentially impossible to obtain neutral outputs.

The EU has a rich diversity of languages. This diversity is not matched by available tools for developing and using NLP. Any future development of algorithms needs to be accompanied by bias measurements, which allow a better understanding of the impact predictions have on decision-making. Only in this way can better, more consistent and less discriminatory decisions become a reality.

Endnotes

[111] FRA (2015), pp. 307–335.

[112] FRA (2020).

REFERENCES

Abid, A., Farooqi, M. and Zou, J. (2021a), '**Large language models associate Muslims with violence**', *Nature Machine Intelligence,* Vol. 3, pp. 461–463.

Abid, A., Farooqi, M. and Zou, J. (2021b), '**Persistent anti-Muslim bias in large language models**', ***AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society***, pp. 298-306.

Akpinar, N. J., De-Arteaga, M. and Chouldechova, A. (2021), 'The effect of differential victim crime reporting on predictive policing systems'*, FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 838–849.

Amnesty International Italy (undated), '**Barometro dell'Odio Project**'.

Article 29 Working Party (2017a), Opinion on some key issues of the Law Enforcement Directive (EU 2016/680), WP 258, Brussels, European Commission Directorate-General Justice and Consumers.

Article 29 Working Party (2017b), Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679, WP251rev.01, Brussels, European Commission Directorate-General Justice, p. 20.

Benbouzid, B. (2019), '**To predict and to manage. Predictive policing in the United States**', *Big Data & Society*, Vol. 6, No. 1, pp. 1–13.

Bennett Moses, L. and Chan, J. (2018), '**Algorithmic prediction in policing: Assumptions, evaluation, and accountability**', *Policing and Society*, Vol. 28, No. 7, pp. 806–822.

Binns R., Veale, M., Van Kleek, M. and Shadbolt, N. (2017), '**Like trainer, like bot? Inheritance of bias in algorithmic content moderation**'.

Birkel, C., Church, D., Hummelsheim-Doss, D., Leitgöb-guzy, N. and Oberwittler, D. (2017), ***Der Deutesche Viktimierungssurvey 2017***, Wiesbaden, Bundeskriminalamt.

Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017), '**Enriching word vectors with subword information**', *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligramma, V. and Kalai, A. (2016), '**Man is to computer programmer as woman is to homemaker? Debiasing word embeddings**', *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4356–4364.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Chen, A., Creel, K., Davis, J. Q., Doumbouya, M., Durmus, E., Ermon, S., Castellon, R., Chatterji, N., Demszky, D., Etchemendy, J., Fei-Fei, Li., Finn, C., Gale, T., Gillespie, L., Goel, K., Donahue, C., Ethayarajh, K., Goodman, N., Grossman, S., Guha, N., Ho, D. E., Hong J., Jurafsky, D., Kalluri, P., Khattab, O., Kumar, A., Hashimoto, T., Hsu, K., Huang, J., Karamcheti, S., Henderson, P., Icard, T., Keeling, G., Hewitt, J., Jain, S., Khani, F., Li, X. L., Mirchandani, S., Ma, T., Malik, A., Mitchell, E., Munyikwa, Z., Koh, P. W., Ladhak, F., Li, X., Krass, M., Krishna, R., Kuditipudi, R., Lee, M., Lee, T., Leskovec, J., Levent, I., Manning, C. D., Nair, S., Narayan, A., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Narayanan, D., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Ren, H., Sadigh, D., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K. and Liang,

P. (2021), '**On the opportunities and risks of foundation models**', workshop on foundation models, Center for Research on Foundation Models, Stanford University, United States, 23–24 August 2021.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), '**Language models are few-shot learners**', *Advances in Neural Information Processing Systems,* Vol. 33, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 6–12 December 2020.

Caliskan, A., Bryson, J. J. and Narayanan, A. (2017), '**Semantics derived automatically from language corpora contain human-like biases**', *Science*, Vol. 356, No. 6334, pp. 183–186.

Charitidis, P., Doropoulos, S., Vologiannidis, S., Papastergiou, I. and Karakeva, S. (2020), '**Towards countering hate speech against journalists on social media**', *Online Social Networks and Media,* Vol. 17.

Chaubard, F., Fang, M., Genthial, G., Mundra, R. and Socher, R. (2019), '**CS224n: Natural language processing with deep learning 1 – Lecture notes: Part I**', course instructor: Manning, C., Winter 2021.

CJEU (Court of Justice of the European Union) (1991), C-184/89, *Helga Nimz v. Freie und Hansestadt Hamburg*, 7 February 1991.

CJEU (1995), C-317/93, *Inge Nolte v. Landesversicherungsanstalt Hannover*, 14 December 1995.

CJEU (1989), C-171/88, *Ingrid Rinner-Kühn v. FWW Spezial-Gebäudereinigung GmbH & Co. KG*, 13 July 1989.

CJEU (1990), C-33/89, *Maria Kowalska v. Freie und Hansestadt Hamburg*, 27 June 1990.

CJEU (1994), C-343/92, *M. A. De Weerd, née Roks, and others v. Bestuur van de Bedrijfsvereniging voor de Gezondheid, Geestelijke en Maatschappelijke Belangen and others*, 24 February 1994.

CJEU (1996), C-13/94, *P v. S and Cornwall County Council*, 30 April 1996.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J. and Liu, R. (2020), '**Plug and play language models: A simple approach to controlled text generation**', *Eighth International Conference on Learning Representations (ICLR 2020),* 26 April–1 May 2020.

Davidson, T., Bhattacharya, D. and Weber, I. (2019), '**Racial bias in hate speech and abusive language detection datasets**', *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), '**BERT: Pre-training of deep bidirectional transformers for language understanding**', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Vol. 1, pp. 4171–4186.

Dixon, L., Li, J., Sorensen, J., Thain, N. and Vasserman, L. (2018), '**Measuring and mitigating unintended bias in text classification**', *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*, pp. 67–73.

Dreißigacker, A. (2017), *Befragung zur Sicherheit und Kriminalität: Kernbefunde der Dunkelfeldstudie 2017 des Landeskriminalamtes Schleswig-Holstein*, Hanover, Kriminologisches Forschungsinstitut Niedersachsen.

Duarte, N., Llansó, E. and Loup, A. (2017), *Mixed messages? The limits of automated social media content analysis,* Washington, D.C., Center for Democracy & Technology.

Dutch Parliamentary Committee (2020), '**Unknown injustice: Report of the Parliamentary Hearing Committee childcare benefits scandal**', 17 December 2020.

ECtHR (European Court of Human Rights), *Boacă and Others v. Romania*, No. 40355/11, 12 January 2016.

ECtHR, *Đorđević v. Croatia*, Merits, No. 41526/10, 24 July 2012.

ECtHR, *M. C. and A. C. v. Romania*, No. 12060/12, 12 April 2016.

ECtHR, *Opuz v.Turkey*, No. 33401/02, 9 June 2009.

ECtHR, *Osman v. the United Kingdom*, Merits, No. 23452/94, 28 October 1998.

Egbert, S. and Krasmann, S. (2019), ***Predictive policing. Eine ethnographische Studie neuer Technologien zur Vorhersage von Straftaten und ihre Folgen für die polizeiliche Praxis***, project completion report, Hamburg, Hamburg University.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. (2018), '**Runaway feedback loops in predictive policing**', *Proceedings of Machine Learning Research*, Vol. 81, pp. 160–171.

European Commission (2018), *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on artificial intelligence for Europe*, COM(2018) 237 final, 25 April 2018.

European Commission (2020), *White Paper On Artificial Intelligence – A European approach to excellence and trust*, COM(2020) 65 final, Brussels, 19 February 2020.

European Commission (2021), ***Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence*** *(Artificial Intelligence Act) and amending certain Union legislative acts*, COM(2021) 206 final, Brussels, 21 April 2021.

European Commission (2022), '**Shaping Europe's digital future – A European approach to artificial intelligence**'.

European Council (2019), *A new strategic agenda 2019–2024*, Brussels, European Council.

European Parliament (2022), Draft report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD)), Brussels, 20 April 2022.

Feltes, T. and Guillen, F. (2020), 'Seguridad y sentimiento de seguridad en Bochum. 40 años de estudio de la cifra negra en una gran ciudad alemana' in: *30 años de la Encuesta de Victimización del Área Metropolitana de Barcelona*, Barcelona, Barcelona Institute of Regional and Metropolitan Studies, pp. 137–153.

Ferris, G., Min, B. and Naya-Oliver, M. (2021), ***Automating Injustice – The use of artificial intelligence and automated decision-making systems in criminal justice in Europe***, London, Fair Trials.

Finck, M. (2019), ***Artificial intelligence and online hate speech***, issue paper, Brussels, Centre on Regulation in Europe.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M. and Kourtellis, N. (2018), 'Large scale crowdsourcing and characterization of Twitter abusive behavior', *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, Vol. 12, No. 1.

FRA (European Union Agency for Fundamental Rights) (2014), ***Violence against women: An EU-wide survey – Main results report***, Luxembourg, Publications Office of the European Union (Publications Office).

FRA (2015), '**Fundamental Rights Agency Survey results – FRA survey in the framework of the eu-LISA pilot on smart borders: Travellers' views on and experiences of smart borders**' in: *Smart Borders Pilot Project – Technical report annexes*, Luxembourg, Publications Office, pp. 307–335.

FRA (2016), ***Ensuring justice for hate crime victims: Professional perspectives***, Luxembourg, Publications Office.

FRA (2017), ***Second European Union Minorities and Discrimination Survey – Main results***, Luxembourg, Publications Office.

FRA (2018a), ***#BigData: Discrimination in data-supported decision making***, Luxembourg, Publications Office.

FRA (2018b), ***Big data, algorithms and discrimination***, Luxembourg, Publications Office.

FRA (2018c), ***Handbook on European non-discrimination law***, Luxembourg, Publications Office.

FRA (2018d), ***Preventing unlawful profiling today and in the future: A guide***, Luxembourg, Publications Office.

FRA (2018e), ***Experiences and perceptions of antisemitism: Second survey on discrimination and hate crime against Jews in the EU***, Luxembourg, Publications Office.

FRA (2018f), ***Being black in the EU***, Luxembourg, Publications Office.

FRA (2019a), ***Data quality and artificial intelligence – Mitigating bias and error to protect fundamental rights,*** Luxembourg, Publications Office.

FRA (2019b), ***Fundamental rights report***, Luxembourg, Publications Office.

FRA (2020), ***Getting the future right – Artificial intelligence and fundamental rights***, Luxembourg, Publications Office.

FRA (2021a), ***Crime, safety and victims' rights***, Luxembourg, Publications Office.

FRA (2021b), ***Your rights matter: Police stops – Fundamental Rights Survey***, Luxembourg, Publications Office.

FRA (2021c), ***Encouraging hate crime reporting – The role of law enforcement and other authorities***, Luxembourg, Publications Office.

Gehman, S., Gururangan, S., Sap, M., Choi, Y. and Smith, N. (2020), '**Real toxicity prompts: Evaluating neural toxic degeneration in language models**', *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369.

German Data Ethics Commission (2019), **Opinion of the Data Ethics Commission**, Berlin, German Data Ethics Commission.

Gershgorn, D. (2018), 'Mark Zuckerberg just gave a timeline for AI to take over detecting internet hate speech', Quartz.

Gerstner, D. (2018), 'Predictive policing in the context of residential burglary: An empirical illustration on the basis of a pilot project in Baden-Württemberg, Germany', *European Journal for Security Research*, Vol. 3, pp. 115–138.

Goldberg, Y. (2017), *Neural network methods for natural language processing*, Canada, Morgan & Claypool.

Gonen, H. and Goldberg, Y. (2019), '**Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them**', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 609–614.

Gorwa, R., Binns, R. and Katzenbach, C. (2020), '**Algorithmic content moderation: Technical and political challenges in the automation of platform governance**', *Big Data & Society*, Vol. 7, No. 1.

Gräber, M. and Horten, B. (2020), '„Werther-Efekt" und „Bullycide"-Medienkonsum, Cybermobbing und Suizidalität von Kindern und Jugendlichen', *Forensische psychiatrie, psychologie, kriminologie*, Vol. 14, No. 4, pp. 467–471.

Greenwald, G. A., McGhee, D. E. and Schwartz, J. L. (1998), 'Measuring individual differences in implicit cognition: The implicit association test', *Journal of Personality and Social Psychology*, Vol. 74, pp. 1464–1480.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M. and Asokan, N. (2018), '**All you need is "love": Evading hate-speech detection**', *AISec'18: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security,* pp. 2–12.

Hardyns, W. and Rummens, A. (2018), 'Predictive policing as a new tool for law enforcement? Recent developments and challenges', *European Journal on Criminal Policy and Research*, Vol. 24, No. 3, pp. 201–218.

Hovy, D. and Spruit, S. L. (2016), '**The social impact of natural language processing**', *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 591–598.

Howard, J. and Ruder, S. (2018), '**Universal language model fine-tuning for text classification**', *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 328–339.

Hunt, P., Saunders, J. and Hollywood, J. S. (2014), *Evaluation of the Shreveport predictive policing experiment*, Santa Monica, RAND Corporation.

Husovec, M. (2021), ***(Ir)responsible legislature? Speech risks under the EU's rules on delegated digital enforcement***, London, London School of Economics.

Johnson, S. D. (2008), 'Repeat burglary victimisation: a tale of two theories', *Journal of Experimental Criminology*, Vol. 4, pp. 215–240.

Katzer, C. (2014), *Cybermobbing*, Cologne, Springer Spektrum, p. 77.

Kennedy, B., Jin, X., Davani, A. M., Dehghani, M. and Ren, X. (2020), '**Contextualizing hate speech classifiers with post-hoc explanation**', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5435–5442.

Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S., Socher, R. and Rajani, N. F. (2020), '**GeDi: Generative discriminator guided sequence generation**', *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, Punta Cana, Association for Computational Linguistics.

Lattacher, S. (2017), '**Predictive policing: Frühwarnsystem für die Polizei**', *Magazin öffentliche Sicherheit*, Vol. 3/4, pp. 11–12.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2020), '**BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pp. 7871–7880.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), '**RoBERTa: A Robustly Optimized BERT Pretraining Approach**', arXiv:1907.11692.

Llansó, E., van Hoboken, J. and Harambam, J. (2020), '**Artificial intelligence, content moderation, and freedom of expression**', Third session of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression, Rockefeller Foundation Bellagio Center, Como, Italy, 12–16 November 2019.

Lum, K. and Isaac, W. (2016), 'To predict and serve?', *Significance Magazine*, Vol. 13, No. 5, pp. 14–19.

Manzini, T., Yao Chong, L., Black, A. W. and Tsvetkov, Y. (2019), '**Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings**', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 615–621.

Marcus, G. and Davis, E. (2021), '**Has AI found a new foundation?**', *The Gradient*, 11 September 2021.

Martini, M. (2021), 'Article 22 GDPR' in: Paal, B. and Pauly, D. (eds.), *DS-GVO BDSG*, 3rd ed., Munich, C.H.Beck.

Mastrobuoni, G. (2020), 'Crime is terribly revealing: Information technology and police productivity', *The Review of Economic Studies*, Vol. 87, No. 6, pp. 2727–2753.

Mittelstadt, B. (2017), 'From individual to group privacy in Big Data analytics, *Philosophy & Technology*, Vol. 30, pp. 475–494.

Mohler, G., Short, M. B., Brantingham, P. J., Schoenberg, F. P. and Tita, G. E. (2011), 'Self-exciting point process modeling of crime', *Journal of the American Statistical Association*, Vol. 106, No. 493, pp. 100–108.

Mohler, G., Short, M., Malinowski, S., Johnson, M., Tita, G., Bertozzi, A. and Brantingham, P. (2015), '**Randomized controlled field trials of predictive policing**', *Journal of the American Statistical Association*, Vol. 110, pp. 1399–1411.

Mosher, C. J., Miethe, T. D. and Hart, T. C. (2010), *The mismeasure of crime*, Thousand Oaks, Sage Publications.

Musto, C., de Gemmis, M., Lops, P. and Semeraro, G. (2021), 'Generating post hoc review-based natural language justifications for recommender systems', *User Modeling and User-Adapted Interaction*, Vol. 31, pp. 629–673.

Nadeem, M., Bethke, A. and Reddy, S. (2021), '**StereoSet: Measuring stereotypical bias in pretrained language models**', *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1, pp. 5356–5371.

OECD (Organisation for Economic Co-operation and Development) (2022), ***OECD framework for the classification of AI systems,*** OECD Digital Economy Papers, No. 323, Paris, OECD Publishing.

OHCHR (Office of the United Nations High Commissioner for Human Rights) (2020), '**One-pager on "incitement to hatred"**'.

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M. and Marco, F. (2020), '**Bias in word embeddings**', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 446–457.

Park, J. H., Shin, J. and Fung, P. (2018), '**Reducing gender bias in abusive language detection**', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2799–2804.

Pennington, J., Socher, R. and Manning, C. D. (2014), '**GloVe: Global Vectors for Word Representation**', *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 1532–1543.

Perry, W., McInnis, B., Price, C., Smith, S. and Hollywood, J. (2013), ***Predictive policing: The role of crime forecasting in law enforcement operations,*** Santa Monica, RAND Corporation.

Perspective (undated), '**Case studies**'.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018), '**Deep contextualized word representations**', *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 2227–2237.

Polizei Berlin (2020), '**Kriminalitätsatlas**'.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018), ***Improving language understanding by generative pre-training***, technical report, OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019), '**Language models are unsupervised multitask learners**', OpenAI Blog, 14 February 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J. (2019), '**Exploring the limits of transfer learning with a unified text-to-text transformer**', *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1-67.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016), '**"Why should I trust you?" Explaining the predictions of any classifier**', *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Richardson, R., Schultz, J. and Crawford, K. (2019), '**Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice**', *NYU Law Review,* Vol. 94, No. 192, p. 218.

Rosen, G. (2022), '**Community standards enforcement report, first quarter 2022**', Meta, 17 May 2022.

Sachdeva, P. S., Barreto, R., von Vacano, C. and Kennedy, C. J. (2022), '**Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus**', FAccT'22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 2022, pp. 1585–1603.

Sanh, V. (2019), '**Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT**', Medium.com, 28 August 2019.

STOA (Scientific Foresight Unit), European Parliamentary Research Service (2020), ***The ethics of artificial intelligence: Issues and initiatives***, PE 634.452, Brussels, March 2020.

Sherman, L., Gartin, P. and Buerger, M. (1989), 'Hot spots of predatory crime: Routine activities and the criminology of place', *Criminology*, Vol. 27, No. 1, pp. 27–56.

Speer, R. (2017), '**ConceptNet Numberbatch 17.04: Better, less-stereotyped word vectors**', 24 April 2017.

Strikwerda, L. (2020), '**Predictive policing: The risks associated with risk assessment**', *Police Journal: Theory, Practice and Principles*, Vol. 94, No. 3, pp. 1–15.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M. and Klenner, M. (2019), '**Overview of GermEval Task 2, 2019 shared task on the identification of offensive language**', *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pp. 354–365.

The Law Society (2019), Algorithms in the criminal justice system, London, The Law Society, p. 36.

Tita, G. and Ridgeway, G. (2007), 'The Impact of gang formation on local patterns of crime', *Journal of Research in Crime and Delinquency*, Vol. 44, No. 2, pp. 208–237.

Townsley, M., Homel, R. and Chaseling, J. (2003), 'Infectious burglaries. A test of the near repeat hypothesis', *British Journal of Criminology*, Vol. 43, pp. 615–633.

Twitter (2022), '**Rules Enforcement**'.

Van Sant, K., Fredheim, R. and Bergmanis-Korāts, G. (2020), 'Abuse of power: Coordinated online harassment of Finnish government ministers', NATO Strategic Communications Centre of Excellence, 24 February 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017), '**Attention is all you need**', *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.

Verma, S. and Rubin, J. (2018), '**Fairness definitions explained**', *Proceedings of the International Workshop on Software Fairness (FairWare)*, pp. 1–7.

von der Leyen, U. (2019), *A Union that strives for more: My agenda for Europe*.

Wachter, S. (2019), 'Affinity profiling and discrimination by association in online behavioural advertising', *Berkeley Technology Law Journal*, Vol. 35, No. 2.

Wachter, S., Mittelstadt, B. and Floridi, L. (2017), 'Why a right to explanation of automated decision-making does not exist in the general data protection regulation, *International Data Privacy Law*, Vol. 7, No. 2, pp. 76–99.

Wachter, S., Mittelstadt, B. and Russell, C. (2021a), 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI', *Computer Law & Security Review*, Vol. 41.

Wachter, S., Mittelstadt, B. and Russell, C. (2021b), 'Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law', *West Virginia Law Review*, Vol. 123, No. 3, p. 25.

Weimann, G. and Masri, N. (2021), '**TikTok's spiral of antisemitism**', *Journalism and Media*, Vol. 2, No. 4, pp. 697–708.

Widinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzedeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G. and Gabriel, I. (2021), '**Ethical and social risks of harm from language models**', Deep Mind.

Wilson, J. and Kelling, G. (1982), '**Broken windows**', *The Atlantic*, March 1982.

Winston, A. (2018), '**Palantir has secretly been using New Orleans to test its predictive policing technology**', *The Verge*, 27 February 2018.

Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M. and Klein, D. (2021), '**Detoxifying language models risks marginalizing minority voices**', *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2390–2397.

Zhao, J., Zhou, Y., Li, Z., Wang, W. and Chang, K.-W. (2018), 'Learning gender-neutral word embeddings', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4847–4853.

ANNEX I: GLOSSARY

Term	Description
Crime observability	A measure of how likely the police would be to observe a crime if present.
Deep learning	A subset of machine learning, involving the use of artificial neural networks (see also 'Neural network') with more than one hidden layer.
Downsampling	Assigning certain probabilities of recording to crime events to counteract the overly strong predictions.
Earthquake policing model	A predictive policing model that relies on criminological research suggesting that crime can spread through local environments through a contagion-like process. This is based on the assumption that, if crime occurs at a certain place, there is a higher likelihood that another crime will take place in the vicinity.
Equalised odds metric	A fairness metric that requires the false positive rates and the false negative rates for a model to be equal across all demographics of interest.
False negative	A type of classification error made by classification models that perform binary labelling, as in an offensive speech detection model that predicts text as either offensive (positive) or non-offensive (negative). If the model falsely identifies a comment as non-offensive, when it is in fact offensive, this is called a false negative.
False positive	See 'False negative'. If the model identifies a comment as offensive despite it being rated as non-offensive in the original or test data, this is called a false positive.
Feedback loop	Decisions based on predictions made by the system influence the data that are then used to retrain or update the system.
Hotspot policing	A strategy that considers an area of a city as a grid of cells and allocates the majority of the police patrols to a fixed number of cells ('hotspots') with the highest predicted risk of crime.
Language model	A system that has been trained to compute the probability of the occurrence of a number of words in a particular sequence. Systems trained in this way on a large body of texts have been found to retain some important semantic features and are at the heart of recent advances in NLP.
Logistic regression	A function often used in binary classification schemes (i.e. classification schemes that have to choose between two possibilities). It basically transforms the question of "what is the probability of x happening" into a linear function (using the log odds function, which takes a probability, p, and turns it into $log[p/(1 - p)]$).
Model	In machine learning, a model is the product of training a machine learning algorithm on training data. A simple example is if linear regression is the algorithm used for predicting the value of variable y based on the values of variable x, then the original function will look like $y = ax + b$. Once the algorithm has been trained, it will have determined the values of a (e.g. 3) and b (e.g. 1), and then the function $y = 3x + 1$ is the model that can then be used to predict y based on new input data x.
Naive Bayes	Naive Bayes classifiers are a family of simple probabilistic classifiers that are based on conditional probability, which is the probability of the occurrence of something based on the occurrence of something else (e.g. the likelihood of a text being offensive given that the text contains the term 'hate').
Natural language processing	The field of designing methods and algorithms that take unstructured, natural language data as the input or produce it as the output. The goal of NLP is to create algorithms that can process natural language, in order to perform a task, ranging from easy (such as spellchecking or keyword searching) to more complex (such as machine translation, question answering or sentiment analysis) tasks.
Neural network	In machine learning, a neural network is a series of interconnected computational units, organised into 'layers', that accept multiple inputs and produce one output. Deep neural networks consist of several layers.
Observed crime	The level of crime observed by the police.
Offensive speech	Speech that may, at a minimum, cause distress to a person. Such speech may incite hatred or threat of violence, which may, at worst, threaten the right to life and physical integrity of people.
Overfitting	A process whereby machine learning is known to potentially focus too strongly on patterns in training data, even when those patterns are random or irrelevant to other situations.

Term	Description
Parameter	In statistics, a parameter is a statistical value (such as the mean or variance) used to describe a statistical population. In computer science, a parameter is a variable whose value needs to be set when calling a function or subroutine of a program.
Predictive policing	The application of analytical techniques – particularly quantitative techniques – to identify likely targets for police intervention and to prevent crime or solve past crimes by making statistical predictions.
Pre-trained language models	AI models that enshrine rules for the usage of words and language based on large corpora of texts (e.g. Wikipedia). Rules of language are pre-trained (i.e. derived from previous texts) and can then be applied and updated with new data for a specific task.
Probabilistic model	Police patrols are distributed according to the crime distribution in historical data. For example, 30 % of patrols are sent to the region where historical records indicate 30 % of crime occurs.
Recorded crime	Observed crime plus reported crime.
Regularisation	A technical solution that should be employed to avoid predictions from becoming too strong. It involves adding an additional parameter to the mathematical formulae of the algorithm.
Reported crime	The level of crime that is reported to police by victims or witnesses.
Runaway feedback loop	When feedback causes a 'winner takes all' situation, for example by repeatedly augmenting the number of police patrols sent to the same neighbourhood regardless of the true crime rate.
Sampling bias	This occurs when some members of a statistical population are systematically more likely to be selected in a sample than others.
Sentiment analysis	The computational study of opinions, sentiments and emotions expressed in text.
Transfer learning	Fully trained machine learning models for word and sentence predictions, which can be used and adapted for new tasks.
True crime rate	In simulation studies, data are artificially created to simulate the reality. One of the parameters included in such a simulation is the assumed 'true crime rate', which may be different from the detected and reported crime rates. It allows for the measure of how much the detected and reported crimes lead to biased predictions compared with the assumed 'true crime rate'.
Unbiased data	Where the historical data are equal to the true crime rates.
Word embeddings	Vector space models of language that represent each word as numbers, where words with similar meanings also have similar numerical values.

ANNEX II: ADDITIONAL TECHNICAL DETAILS OF THE SIMULATION OF FEEDBACK LOOPS

Predictive policing is a process with several actors and possible sources of biases. Consider, for example, that certain types of crime are reported less than others,[113] as is the case for crime incidents with differential observability by police patrols.[114] In general, the complexity of human behaviour and cultural socio-economic biases permeate into historical datasets, making it difficult to keep track of all the factors influencing the process. Since the goal is to investigate the formation of feedback loops, some of the real-life complexities can be avoided, and a more simplified approach can be taken by focusing on the following parameters only. Crime reporting behaviour is represented in a single parameter. The factors that influence the detection of crime include the police patrol distribution (β) and the observability of crime (V). Finally, a fourth parameter is needed to reflect all crime events happening in a given city. Figure 2 describes the components of the simulated policing process and how they are connected to one another (through the directed arrows), and the following list provides detailed descriptions of the parameters.

- **Parameter α: crime reporting rates.** These concern the reports submitted by witnesses or victims to the police. It can differ across neighbourhoods, and was held constant during each of the simulations. The value of α is based on data from German and Spanish victimisation surveys[115] and set to 20 % (it was 22.3 % in Barcelona in 2019).
- **Parameter β: police distribution.** This represents the distribution of police patrols through the districts and the related direct observations by the police patrols. Here, it is assumed that the number of observations is directly related to the presence of police patrols per neighbourhood. The initial parameter value was based on the recorded crime rates in the cities (Vienna, Berlin, Madrid and Barcelona) and was set by the developers at the beginning of the simulation. The initial value was based on 2019 car theft data from Berlin.[116] This parameter changed through the evolution of the simulation based on the respective predictions. This is the most relevant parameter for identifying the formation of (runaway) feedback loops. In fact, the system bias is measured by looking at the difference between the true crime distribution and the value of β. The Kullback–Leibler divergence, a common measure of the difference between two distributions, was used to measure this difference.
- **Parameter V: observability of crime.** This is a measure of how likely the police are to detect a crime in a particular neighbourhood. Observability can depend on, for example, the type of crime that is committed, with some crimes being more likely to be detected (e.g. car theft) and others less likely (e.g. tax fraud). For experiment 1, we allow parameter V to depend on crime type and on district, in order to detect the impact of crime observability on feedback loop formation. In this context, crime observability is taken to mean the following: if crime type A has an observability of 30 % in district 1, then, if a crime of type A occurs in district 1, and the police happen to be patrolling there, there is a 30 % chance that they will observe and record the crime. For experiment 2, in order to simplify simulations, crime type is fixed, and observability is assumed to be the same across all districts, so that crime observation depends exclusively on the distribution of patrols.

- **Parameter Ω: true crime distribution.** Compared with the analysis of a real process, the main advantage of a simulation is that we can control parameters that are often unknown. In this case, the true distribution of crime across neighbourhoods is a very difficult parameter to infer from the statistics of police records. It is well documented that a proportion of crime events never enter police records (the 'dark figure') because they are neither observed nor reported.[117] In addition, several other factors influence those statistics, such as low rates of victims reporting crime,[118] limited police resources or the so-called grey zone,[119] which corresponds to crimes that are reported to police but are not recorded in the crime statistics. Nonetheless, the simulations in this research explore a large set of true crime rate distributions, Ω, to study the influence of the true crime rate distribution on the formation of feedback loops.

- **Other parameters.** For the implementation of the simulation, setting some standard values is required. It is assumed that there are, on average, 200 crime events of a certain type per month. This is, for example, the approximate number of highway robbery events in Berlin in 2019, the illegal use of drugs in Madrid recorded in April 2019 and the sum of mugging events in two neighbourhoods in Barcelona in June 2019. Additional parameters include the duration of the simulation (number of epochs, which was equivalent to 20 and 100 years); the length of historical data (duration of initial data: approximately 365 days) and the algorithm update frequency (duration of each epoch: approximately seven days), equivalent to one week. In addition, experiment 1 had two settings. The first kept all historical data, thus accumulating an ever larger dataset as the simulation progressed through time. The second always kept only one year's worth of data (thus, 'forgetting' everything that was older than one year). This second setting was also used in experiment 2.

Table A1 shows the sources of crime data used to specify values of parameters.

TABLE A1: CRIME DATA USED TO INFORM PARAMETER ASSUMPTIONS

City	Number of districts	Period	Number of types of crime	Source
Vienna	23	2017–2019	> 200	Federal Criminal Office Austria
Berlin	12 (138 subdistricts)	2012–2020	16	Kriminalitätsatlas (Berlin.de)
Madrid	21	2014–2020	27	Madrid Open Data Portal police statistics
Barcelona	9	2011–2020	23	Policia de la Generalitat de Catalunya

POLICING STRATEGIES

An assumed policing strategy is the part of the simulation where a decision based on the output from the predictive policing algorithm is modelled. It is part of the process of allocating police according to predictions, as shown in Figure 2.

Two policing strategies are considered.

- **Allocating police patrols proportionally to the predictions made by the algorithm.** For example, if the algorithm predicts 20 % of crime events in district 1, 20 % of police patrols are assigned to district 1. This strategy was used mainly in simulation 1, described below, and is sometimes referred to as 'effective policing'.
- **Hotspot policing.** This considers a city as a grid of cells (e.g. of 5 × 5 m^2 each), and allocates the majority of police patrols to the hotspots – that is, to a fixed number of cells (n) with the highest probability of crime. For example, if $n = 5$ and there is a hypothetical city with 25 cells, this strategy selects the five cells with the highest probability of crime and dispenses the majority of police patrols to those five cells. This strategy was mainly used in simulation 2, described below.

PREDICTION ALGORITHMS

Various algorithms were tested. The algorithms and models include:

- a simple probabilistic model, which allocates police patrols according to the recorded crime rate, including reported and detected crime, in each neighbourhood and is not based on a machine learning model (used in simulation 1);
- simple machine learning algorithms, including so-called naive Bayes and logistic regression (used in simulation 1);
- a more complex model, referred to as the 'earthquake policing model', which is described in more detail below.

NUMBER OF NEIGHBOURHOODS

Simulation 1 explores the case of two neighbourhoods. Using only two neighbourhoods makes visualisations easier to understand and the effects of various parameters more clearly observable. However, simulation 1 was also conducted for more than two neighbourhoods, and runaway feedback loops were also formed in those situations. Simulation 2 experiments were all conducted for multiple neighbourhoods.

Simulation 1: Two neighbourhoods, simple statistical and machine learning models, effective policing
The first simulation involves only two neighbourhoods. Initially, a general probabilistic model is used. This model takes a simple statistical approach and assumes that police patrols should be distributed exactly according to historical crime records. For example, if police records indicate that 70 % of crime happens in neighbourhood 1, and 30 % in neighbourhood 2, then 70 % and 30 % of patrols should be sent to neighbourhoods 1 and 2, respectively (i.e. effective policing). This is used as a basic model to explore the interplay of the parameters and to broadly understand the predictive policing process. The main analysis was carried out using a simple machine learning algorithm (called naive Bayes), trained on historical crime records, which predicts the percentage of crime across two neighbourhoods on the following day. The results of the probabilistic model were compared with the predictions of the naive Bayes model. This enables understanding of the effect of using the different algorithms, and observes how much the feedback loop formation process is affected by using machine learning algorithms compared with the first simple allocation according to the proportion of recorded crime for each district. In addition, to corroborate the validity of the results, the performance of the naive Bayes model was compared with that of a logistic regression model, another commonly used, but relatively simple, machine learning algorithm. It is important to note that the formation of feedback loops in this context is an artefact of the system design (because the system affects the data that are then used to update the same system) and is independent of the presence or absence of a machine learning model for prediction. Even in the simple probabilistic setting, runaway feedback loops are formed, as shown in previous research.[120] The simulations show, however, that the inclusion of machine learning models accelerates the formation of the runaway feedback loops, at least in settings where no mitigation measures are implemented.

Simulation 2: Multiple neighbourhoods, more complex machine learning models, hotspot policing
Simulation 2 uses the hotspot policing strategy and inspects its performance in several neighbourhoods, including an example in which 2019 Berlin crime records are used. The algorithm used is a special case of a so-called self-exciting point process. It was first introduced to predictive policing by a team of researchers that included the cofounder of PredPol, George Mohler, in 2011.[121] This is often referred to as the 'PredPol model' in the literature. However, it is not clear to what extent the current commercial application corresponds to the original theoretical model described in the literature. Therefore, the model is referred to as the 'earthquake policing model' in this report.[122] The rationale behind the use of the earthquake policing model for predictive policing relies on criminological research suggesting that crime can spread through local environments through a contagion-like process,[123] particularly for certain types of crimes such as vandalism to property,[124] burglary and gang violence.[125] However, the use of the earthquake policing model in the area of predictive policing has been controversial since its inception.[126] This is partly due to the underlying assumptions of the process and some incompatibilities between the nature of the earthquake policing model and how crime emerges in cities. It has also been controversial because of the tendency of the model to form feedback loops.[127]

In analogy with the published earthquake policing model,[128] hotspot policing is assumed in simulation 2. This means sending a majority of police to the *n* cells with the highest predicted risk of crime (the crime hotspots). We use a ratio of 5:1, meaning five times as many police were assumed to be sent to a crime hotspot than to cells with a lower predicted risk of crime. The earthquake policing model is used to predict the *n* cells with a higher risk of crime. In the frame of simulation 2, four main experiments to study the formation of feedback loops were run.

- The first experiment uses a grid of 5 × 5 m^2 cells, with equal true crime rates across the cells and initial historical rates distributed uniformly with the exception of five cells, which have twice the rates of the other cells to emulate regions with higher crime rates. The simulation is run over 700 iterations ('days'), which means there were 700 updates when a new police assignment to visit the five cells with the highest crime rates was made.
- The second experiment considers a **non-uniform distribution of true crime rates**. To set up a realistic situation, but one that is still simple to interpret, we use crime statistics on car thefts in Berlin (2019). Furthermore, in order to extract the ratio of the top 30 % of subdistricts with respect to the other cells, we first estimate the average rates within both (top 30 % and the remaining 70 %) groups, and calculate the division. A ratio of about 1:24 is obtained from these data. In our simulations of a grid of 5 × 5 m^2 cells, this means setting eight cells (approximately the top 30 %) with rates higher than the rest by about 25 %. In that way, the ratio in our simulations is 1:25.
- The third experiment also considers car theft events for a grid of 5 × 5 m^2 cells with **uniform true crime rates** across the cells. The total number of areas in the dataset is 138. Thus, we randomly select 25 rates for the data

to be used in the experiment. We observe the simulation over 1,000 iterations, which is 1,000 simulated ‘days’ in the computer simulation.

- The fourth experiment is similar to the third. It extracts the true crime rates from the dataset, which is equivalent to assuming that the data are unbiased. This is probably the best setting to observing whether feedback loops form and how fast the process is. The simulation is run over 1,000 iterations (‘days’).
- Finally, the fourth experiment was also run on the full dataset, based on 138 subdistricts of Berlin. The data are distributed on a grid of 23 × 6 m^2 cells, with historical and true crime rates corresponding to 2019 Berlin car theft statistics. The simulation is run over 1,000 iterations (‘days’), and the police assignment will be distributed across the 20 cells with highest predicted risk of crime.

ANNEX III: TECHNICAL DETAILS OF THE OFFENSIVE SPEECH DETECTION ANALYSIS

OFFENSIVE SPEECH DETECTION ALGORITHMS

NLP is the field of designing methods and algorithms that take as input or produce as output unstructured, natural language data.[129] The goal of NLP is to create algorithms that can process natural language to perform a particular task, ranging from easy (such as spellchecking or keyword searching) to more complex (such as machine translation, question answering or sentiment analysis) tasks.[130] Early statistical models for NLP were based on very simple supervised learning techniques. Text is decomposed into its basic units – words – without care for word order. This choice is by no means neutral and is indicative of an English-language bias. In languages that are morphologically richer than English, words more often have different meanings depending on the context and grammar. Words may not be the ideal basic unit of meaning in such languages.[131] This approach of using words as input without considering their order is called the bag-of-words approach. These 'bags of words' are then used as the 'input features' or predictors for the task at hand.

Word embeddings

Word embedding is a method of mapping words onto a numerical vector space. Put more simply, this means it transforms words into several numbers, while still preserving certain important semantic relationships. This means that words with similar meanings are also closer together (i.e. more similar) in their numerical representation.

Once the word embeddings have been established ('learned'), they can be saved and reused in many other NLP tasks. The use of these pre-trained word embeddings overcomes the limitation of the traditional approach, in which classifiers only learned how to deal with words already encountered in the training set. Another strong appeal of the word embeddings method is that it is unsupervised. It requires a lot of text data, but no labelling, as the 'closeness' of words is drawn from existing text. Thus, the pre-trained word embeddings can encompass ever larger vocabularies. Furthermore, the embeddings preserve some form of semantic relationship. This refers, in essence, to the probability of a given word belonging in a particular context, 'learned' from the vast corpora used during training of the word embedding. This allows subsequent models to take advantage of the semantic relationships already encoded in the word representations.

In general, huge bodies of text datasets were required to train the word embeddings, and these were taken from the internet. Wikipedia and selected social media platforms such as Twitter and Reddit are among the most popular sources. However, it soon transpired that not all associations learned from such text datasets were desirable. The first research paper to reach public consciousness and raise the alarm was entitled 'Man is to computer programmer as woman is to homemaker? Debiasing word embeddings'.[132] This demonstrated that word embeddings preserved gender-biased associations, and proposed a mathematical procedure for debiasing the word vector space. Using similar techniques, the article 'Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings'[133] tackled the issue of detecting and mitigating racial bias in word embeddings. Caliskan *et al.*[134] followed a different strategy. They developed the word embedding association test, based on the implicit association test from psychology.[135] With the help of this test, they were able to detect a wide variety of human biases in word embeddings trained on large public corpora. Finally, using the word embedding association test to conduct experiments on some gender-debiased embeddings,[136] Gonen and Goldberg[137] point out that debiasing techniques could be merely hiding the bias, rather than removing it.

Much of the research on bias in word embeddings is based on English. One study on bias in German word embeddings[138] finds that debiasing methods developed in English are not appropriate for gendered languages, and suggests that a debiasing method for German should be developed. Unfortunately, no pre-trained debiased German word vectors were provided. For this research, only the ConceptNet[139] word embeddings for pre-trained and debiased word vectors in languages other than English could be found.

Language models

The early generation of word embeddings, as described above, were context independent (i.e. each word has exactly one embedding, regardless of context). The subsequent generation of word embeddings, such as GPT,[140] ELMo[141] and ULMFiT,[142] produced embeddings of words in context, so that the vector representation of a word depended also on its neighbouring words. Shortly thereafter, starting in 2019, large-scale pretrained language models were developed, such as BERT,[143] GPT-2,[144] GPT-3,[145] RoBERTa,[146] T5[147] and BART.[148] The distinguishing feature of these language models is that they are fully trained machine learning models. Unlike the pre-trained word embeddings, which used machine learning to produce word vectors that could be used for downstream tasks, the model itself is used in a downstream task (after only some initial 'fine-tuning' per task). An example of this is known as 'transfer learning'. This process is based on two novelties. The first is that the unsupervised task used to pre-train the models usually involves not only word predictions, but also next-sentence prediction. The second, and more significant, is that a new underlying neural network architecture – transformers – was used.[149] These language models have been competing with each other in terms of sheer size of vocabularies used in training, and in terms of the number of parameters included in their feature space.[150] Their promise is that, after the initial expense of developing ('training') such models, they can be used as the foundation for other algorithms, which can fine-tune them and be adopted to accomplish downstream tasks with limited amounts of labelled training data. In fact, with this potential in mind, over 100 researchers convened at Stanford University in August 2021 to announce a new paradigm in the field of machine learning: foundation models.[151] This development was met with some scepticism. As noted by Marcus and Davis,[152] "Large pretrained statistical models can do almost anything, at least enough for a proof of concept, but there is precious little that they can do reliably".

Furthermore, evidence is mounting that language models are subject to the same issues of bias as word embeddings. Research indicates that gender, racial[153] and religious[154] stereotypes are embedded in the pretrained language models. Methods to properly measure this bias are being investigated, since the tools developed for word embeddings do not necessarily adapt well to language models.[155] While there have been attempts to develop strategies to remove the bias in language models,[156] research indicates that such debiasing efforts might have negative consequences for marginalised voices.[157] The negative effect could be that debiasing could lead to diminished model performance on language used by minority groups. The potential for bias is particularly troubling in the context of pre-trained language models as building blocks in countless other NLP models for a wide range of tasks. Without a better understanding of how the bias propagates into downstream tasks, and effective measures for mitigating it, there is a serious risk of perpetuating and exacerbating biases and discrimination picked up by the pretrained language models.

ALGORITHMS USED FOR DEVELOPING THE MODELS

Models based on three different machine learning algorithms were developed using publicly available labelled datasets in English,[158] German[159] and Italian,[160] as described above.

- **Model 1.** This is a simple bag-of-words approach using logistic regression, as in Davidson *et al.*[161]
- **Model 2.** This is a deep neural network and word embeddings approach. For this algorithm, two initial choices had to be made: the particular word embeddings and the choice of the neural network architecture. As for word embeddings, an initial investigation into GloVe[162] showed that no pre-trained embeddings of comparable size and quality exist in English, German and Italian. While the embeddings that were used (**fastText**[163]) were available in all three languages, the English-language embeddings were unavoidably larger. However, all embeddings were trained within the same framework to ensure a higher chance of consistent quality. fastText also makes the claim that it is better suited to morphologically more complex languages, by training not just on words, but combining subword information in its algorithm training procedure.[164] For the algorithm, convolutional neural network and gated recurrent unit architectures were tested. To allow for comparison, the baseline architectures were taken from Park *et al.*[165] and the **Conversationai GitHub**.
- **Model 3.** To train an offensive speech detection model using pre-trained models, we used the **Hugging Face Transformers** library. This is an open-source library, which contains implementations of many of the state-of-the-art language models. The DistilBERT[166] implementation was used during development runs. In the end, the development used the BERT[167] and RoBERTa[168] multilingual versions. In all three languages, the BERT-based models performed the best.

Each algorithm was trained separately on each of the three datasets, giving a total of nine models to be used in testing for bias.

Algorithm training involves several choices that need to be made. As indicated above, several alternatives were investigated in order to find the best results. For example, for model 2, a choice needed to be made between convolutional neural network or gated recurrent unit neural network architectures. Other details needed to be decided experimentally, and therefore it is common practice to split data into the following three sets.

- **Training dataset.** This is used to train each algorithm, and obtain a model.
- **Validation dataset.** This is used to test the trained model performance, in order to choose the 'best' one.
- **Test dataset.** Once a 'best' model has been identified, it is retrained on a combined dataset of training and validation data. Its performance is then tested on the test data, and this is what is reported when giving the model performance.

For the bias analysis, an additional dataset was used for testing for bias. This dataset was generated from templates to test for bias against various identities. It includes invented text phrases that were used to obtain predictions of offensiveness. This dataset is referred to as the 'bias test dataset', and is described in more detail below.

DATASETS (TRAINING DATA) USED TO BUILD OFFENSIVE SPEECH DETECTION ALGORITHMS

Table A2 provides an overview of the data used for the offensive speech detection model training conducted for this research, which were all based on actual social media comments. English- and Italian-language models were trained on data from one dataset only. Owing to the limited dataset size and quality, two datasets were combined in order to train the German-language models. Before commencing the model training, all datasets were reduced to one field containing the comment and one field containing the offensiveness label. No other variables were used in the model development process. Amnesty International Italy kindly granted permission for the use of the Italian-language dataset.[169]

TABLE A2: CHARACTERISTICS OF TRAINING DATA

Characteristics	Language			
	en	it	de	
Dataset name	Founta *et al.*[170]	Barometro[171]	GermEval[172]	Zenodo[173]
Total number of samples/ observations	99,996	132,868	15,418	27,216
Number of samples after processing	91,633	108,399	15,410	26,953
Offensive (%)	27	8	33	2
Used samples	Non-offensive: 67,006 (73 %) Offensive: 24,627 (27 %)	Non-offensive: 30,177 (75 %) Offensive: 10,059 (25 %)	Non-offensive: 10,322 (67 %) Offensive: 5,088 (33 %)	Non-offensive: 26,328 (98 %) Offensive: 625 (2 %)
Platform	Twitter	Facebook (55 %) and Twitter (45 %)	Twitter	Twitter
Collection period	2018	2018–2020	2018–2019	2018–2019
Context of data collection	Development of automated offensive language detection algorithms	Treatment of women and LGBT people; European Parliamentary elections	Identification of offensive language	Data collection on hate speech in journalism
Associated institution	Aristotle University of Thessaloniki, Cyprus University of Technology, Telefónica Research, University of Alabama at Birmingham, University College London	Amnesty International Italy	Saarland University, Darmstadt University of Applied Sciences, Leibniz-Institut für Deutsche Sprache	DataScouting, International Hellenic University

Characteristics	Language			
	en	it	de	
Category labels	Non-offensive: normal, spam Offensive: abusive, hateful	Non-offensive: neutral/positive; negative, non-problematic Offensive: negative, problematic; hate speech Not included: ambiguous	Non-offensive: ‘Other’ Offensive: ‘Offence’	Non-offensive: ‘Non hate speech’ Offensive: ‘Hate speech’
Labelling strategy	The labels were obtained through crowdsourcing	Offensiveness was scored by trained Amnesty International Italy volunteers	Manually annotated by one of the researchers	The data follow an automated sampling procedure before they are manual labelled

Bias test dataset

A special bias test dataset was generated, based on the templates developed by Dixon *et al*.[174] The sentences contain offensive and non-offensive phrases, based on the negative/positive adjectives (negative adjectives: disgusting, dumb, useless, primitive, etc.; positive adjectives: great, smart, impressive, strong, etc.) or verbs (negative verbs: hate, kill, destroy, abuse, etc.; positive verbs: like, love, respect, support, etc.) that were used in the sentences. Various identity terms were developed and used (e.g. German, Irish, Muslim, migrant) to test for biases across different protected characteristics. These are indicated in the ‘Collective noun’ columns in Table A3. The missing words were filled in from lists in order to generate an approximately equal number of offensive and negative sentences for each identity term. The same templates and word lists were translated into all three languages. The German and Italian sentences also used gendered terms. Overall, this led to a dataset of over 7,300 sentences, in which about half of the sentences were rated by the research team as offensive and the other half were rated as non-offensive. The German- and Italian-language bias test datasets were twice as large as the English-language one, because the terms were gendered. Each of the example sentences in each language was predicted for its offensiveness in each of three models. This led to over 110,000 predictions, which were investigated for bias according to the identity terms used.

Table A3 provides examples of sentences developed for testing bias in the models. The sentences were largely kept the same, with only certain selected terms being changed.

TABLE A3: EXAMPLE SENTENCES FOR THE BIAS TEST DATASET

EXAMPLE 1

Language	Verb (positive/negative)		Article (feminine/masculine)	Collective noun	Sentence
English	abuse	all		Whites	Abuse all whites.
German	*missbrauche*	*alle*		*Weißen*	*Missbrauche alle Weißen.*
Italian	*abusare*	*di tutte/tutti*	*i*	*Bianchi*	*Abusare di tutti i bianchi.*

EXAMPLE 2

Language		Article (feminine/masculine)	Collective noun		Sentence
English	I think that all		males	should be killed	I think that all males should be killed.
German	*Ich denke, dass alle*		*Männer*	*umgebracht werden sollen*	*Ich denke, dass alle Männer umgebracht werden sollen.*
Italian	*Penso che tutti*	*i*	*Maschi*	*dovrebbero essere uccisi*	*Penso che tutti i maschi dovrebbero essere uccisi.*

EXAMPLE 3

Language		Article (feminine/ masculine)	Adjective (positive/ negative, feminine/ masculine)	Collective noun	Sentence
English	You are	a	smart	Buddhist	You are a smart Buddhist.
German	*Du bist*	*ein*	*kluger*	*Buddhist*	*Du bist ein kluger Buddhist.*
			Collective adjective	**Adjective (positive/ negative, feminine/ masculine)**	
Italian	*Sei*	*un*	*buddista*	*intelligente*	*Sei un buddista intelligente.*

PERFORMANCE OF ALGORITHMS

This section provides an overview of the overall performance of algorithms in terms of accurately detecting offensive comments. After the three models were trained in each of the three languages (English, German and Italian), their performance was evaluated on the test datasets. The so-called F1-score was used to capture how well the offensive comments were being predicted. The F1-score is a metric of model performance for classification models that have to identify only two classes ('positive' or 'negative'). It is often used when trying to measure how well a model detects the 'positive' class, especially when the 'positive' class makes up a smaller proportion of the test data. The results are summarised in Table A4, alongside several commonly used performance metrics for each of the models, on different datasets.

All three English-language models significantly outperformed the corresponding models in German and Italian. This is true for the test set performance, and for the bias test set performance.

All languages and all models showed a big drop in performance between their results on the test set and their results on the bias test set.

The test set, as described above, while not used in the training of the models, was sampled from the same data as the training data, and thus has the same statistical and linguistic properties.

The bias test set, on the other hand, has different statistical properties, and this led to a severe drop in performance.

In English, even the simple model performed very well on the test set, and the extra computing power and resources used to train the neural networks solution (model 2) and the language model (model 3) resulted in only marginal improvements. This raises the question of whether the marginal gain in performance is worth the extra cost and complexity of developing models 2 and 3.

Simple algorithms, such as the logistic regression used for model 1, are fully determined once the training data and some initial parameters are known. This means that, given the same data, and the same initial parameters, the same logistic regression model will always be built. This is not the case for more complex models, which often involve several random starting points. There are methods for ensuring that results can be reproduced when the same data, the same parameters and the same system set-up are used. Unfortunately, neural networks can be so complex that this is not guaranteed. To check the reproducibility of the experiment results, once the optimal architecture had been obtained, models 2 and 3 were retrained 10 times, and each of the 10 versions of each model was used to evaluate all the datasets. Model 3 versions were stable, and all reproduced the original model. Model 2 results varied to some extent in English and Italian, meaning there was some amount of reproducibility. However, the German-language model 2 versions varied significantly, and suffered a great drop in performance. The reasons for this failure in reproducibility could not be established in the frame of this research (see the box 'Challenges and limitations encountered when researching bias in speech detection' in Chapter 3).

TABLE A4: PERFORMANCE METRICS OF MODELS ON TEST DATASETS

Language	Model	Dataset	Performance metric				
			Precision	Recall	F1-score	Accuracy	AUC
en	1	Test set	0.87	0.88	0.91	0.93	0.96
en	2	Test set	0.89	0.88	0.92	0.94	0.97
en	3	Test set	0.89	0.89	0.92	0.94	0.97
en	1	Bias test set	0.76	0.57	0.69	0.70	0.80
en	2	Bias test set	0.81	0.54	0.70	0.71	0.83
en	3	Bias test set	0.85	0.64	0.76	0.76	0.87
de	1	Test set	0.67	0.64	0.73	0.76	0.81
de	2	Test set	0.41	0.88	0.44	0.48	0.64
de	3	Test set	0.77	0.76	0.82	0.83	0.90
de	1	Bias test set	0.66	0.57	0.64	0.64	0.69
de	2	Bias test set	0.54	0.75	0.44	0.53	0.56
de	3	Bias test set	0.73	0.57	0.68	0.68	0.75
it	1	Test set	0.53	0.75	0.73	0.77	0.84
it	2	Test set	0.52	0.70	0.71	0.76	0.82
it	3	Test set	0.72	0.63	0.79	0.85	0.89
it	1	Bias test set	0.59	0.64	0.60	0.60	0.64
it	2	Bias test set	0.58	0.79	0.60	0.61	0.65
it	3	Bias test set	0.68	0.70	0.68	0.68	0.74

Notes: *Precision is the proportion of offensive text predicted correctly within all instances predicted to be offensive. Recall is the proportion of offensive text predicted correctly within all instances of text originally rated as offensive. F1-score combines precision and recall through calculating its harmonic mean. Accuracy is the percentage of predictions that match the labels across all observations, irrespective of whether they are labelled as offensive or non-offensive. AUC, area under the curve. This is another accuracy metric that takes into account all possible thresholds for making a decision on whether text is offensive or not, considering the trade-off between the true positive rate and the true negative rate.*

Endnotes

113 FRA (2021a).
114 Lum and Isaac (2016).
115 Birkel *et al*. (2017); and Murrià *et al*. (2020). In 2019, the average reporting rate was around 20 %.
116 Taken from Polizei Berlin (2020).
117 Dreißigacker (2017).
118 FRA (2021a).
119 Feltes and Guillen (2020).
120 Ensign *et al*. (2018).
121 Mohler *et al*. (2011).
122 Lum and Isaac (2016); and Ensign *et al*. (2018).
123 Johnson (2008).
124 Wilson and Kelling (1982).
125 Tita and Ridgeway (2007).
126 Benbouzid (2019).
127 Lum and Isaac (2016).
128 Mohler *et al*. (2015).
129 Goldberg (2017).
130 Chaubard *et al*. (2019).
131 Hovy and Spruit (2016); and Bojanowski *et al*. (2017).
132 Bolukbasi *et al*. (2016).
133 Manzini *et al*. (2019).
134 Caliskan *et al*. (2017).
135 Greenwald *et al*. (1998).
136 Bolukbasi *et al*. (2016); and Zhao *et al*. (2018).
137 Gonen and Goldberg (2019).
138 Papakyriakopoulos *et al*. (2020).
139 Speer (2017).
140 Radford *et al*. (2018).
141 Peters *et al*. (2018).
142 Howard and Ruder (2018).
143 Devlin *et al*. (2019).
144 Radford *et al*. (2019).
145 Brown *et al*. (2020).
146 Liu *et al*. (2019).
147 Raffel *et al*. (2019).
148 Lewis *et al*. (2020).
149 Vaswani *et al*. (2017).
150 Sanh (2019).
151 Bommasani *et al*. (2021).
152 Marcus and Davis (2021).
153 Gehman *et al*. (2020).
154 Abid *et al*. (2021a); see also Abid *et al*. (2021b).
155 Nadeem *et al*. (2021).
156 Krause *et al*. (2020).
157 Xu *et al*. (2021).
158 Founta *et al*. (2018).
159 Struß *et al*. (2019); and Charitidis *et al*. (2020).
160 Amnesty International Italy (undated).
161 Davidson *et al*. (2019).
162 Pennington (2014).
163 Bojanowski (2017).
164 *Ibid*.
165 Park *et al*. (2018).
166 Sanh *et al*. (2019).
167 Devlin *et al*. (2019). Official Hugging Face implementations exist for English and German, but there is only a user upload for Italian.
168 Liu *et al*. (2019).
169 Amnesty International Italy (undated).
170 Founta *et al*. (2018).
171 Amnesty International Italy (undated).
172 Struß *et al*. (2019).
173 Charitidis *et al*. (2020).
174 Dixon *et al*. (2018).

Getting in touch with the EU

In person
All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: **european-union.europa.eu/contact-eu/meet-us_en**

On the phone or by email
Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696 or
- by email via: **european-union.europa.eu/contact-eu/write-us_en**

Finding information about the EU

Online
Information about the European Union in all the official languages of the EU is available on the Europa website at: **https:// europa.eu/european-union/index_en**

EU publications
You can download or order free and priced EU publications at: **https://op.europa.eu/en/publications**. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see **european-union.europa.eu/contact-eu/meet-us_en**).

EU law and related documents
For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR- Lex at: **http://eur-lex.europa.eu**

Open data from the EU
The EU Open Data Portal (**data.europa.eu/en**) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

FRA EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS

PROMOTING AND PROTECTING YOUR FUNDAMENTAL RIGHTS ACROSS THE EU

Artificial intelligence is everywhere and affects everyone – from deciding what content people see on their social media feeds to determining who will receive state benefits. AI technologies are typically based on algorithms that make predictions to support or even fully automate decision-making.

This report looks at the use of artificial intelligence in predictive policing and offensive speech detection. It demonstrates how bias in algorithms appears, can amplify over time and affect people's lives, potentially leading to discrimination. It corroborates the need for more comprehensive and thorough assessments of algorithms in terms of bias before such algorithms are used for decision-making that can have an impact on people.

EU Charter of Fundamental Rights

Non-discrimination

Information society

FRA – EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS
Schwarzenbergplatz 11 – 1040 Vienna – Austria
TEL. +43 158030-0 – FAX +43 158030-699

fra.europa.eu

facebook.com/fundamentalrights
twitter.com/EURightsAgency
linkedin.com/company/eu-fundamental-rights-agency

Publications Office
of the European Union