User Study on the Validity of Automatically Generated Legal Texts - General Questions

Welcome to our "User Study on the Validity of Automatically Generated Legal Texts" a web-based experiment designed to study the validity of human-written legal texts versus automatically generated legal texts. Before taking part in this study, please read the consent form below and choose "I Agree" at the bottom of the page if you understand the statements and freely consent to participate in the study.

* Indicates required question

Consent Form

This study is aimed at legal practitioners (e.g., lawyers and law students) as well as computer science professionals (e.g., software developers). It involves monetary compensation for participants.

Our study investigates the validity of the inputs generated by our automated fairness test generation approach for AI applications. We have proposed a semantics-aware mutation-based fairness testing method for NLP systems. Specifically, we have evaluated our approach using several AI-based NLP systems designed for legal datasets such as ECtHR, EUR-Lex, SCOTUS and LEDGAR. The aim of the study is to validate if the discriminatory inputs (i.e., legal texts) generated by our approach are valid and consistent, in comparison to human-written texts. In particular, we examine the grammatical correctness, legal validity, similarity and consistency of the legal text inputs generated by our approach. We explore the expectations of legal professionals and software developers for the legal text generated by our approach.

Participation typically takes 30-60 minutes and the study is strictly anonymous. Participants begin by answering simple questions about their professional background and skill as legal practitioners or software developers. This is followed by a series of questions about the validity, similarity, and consistency of generated legal texts versus human-written legal texts.

All responses are treated as confidential, and in no case will responses from individual participants be identified. Rather, all data will be pooled and published in aggregate form only. All data is stored in a password-protected electronic format. To help protect your confidentiality, the surveys will not contain information that will personally identify you. The results of this study will be used for scholarly purposes only and may be shared with representatives of the XXX.

Many individuals find participation in this study enjoyable, and no adverse reactions have been reported thus far. If participants have further questions about this study or their rights, or if they wish to lodge a complaint or concern, they may contact the project lead, XXX at XXX, or XXX at XXX.

If you are **18 years** of age or older, understand the statements above, and freely consent to participate in the study, click on the "**I Agree"** button to begin the experiment.

1. Please give your consent below: *

   I certify that I acknowledge the above information and consent.

   Check all that apply.

   □ I agree (I hereby consent)

2. Provide your **User ID** (MTurk, Prolific), or your **Study Referral Code,** if any

____

Demographics and Experience

General information about the participant

3. Is English your native language ? *

Mark only one oval.

○ Yes

○ No

4. Which of the following best describes your level of English proficiency? *

Mark only one oval.

○ Novice

○ Intermediate

○ Advanced

○ Expert

5. Rate your level of English proficiency *

Mark only one oval.

	0	1	2	3	4	5	
No k	○	○	○	○	○	○	Very proficient (e.g., native speaker)

6. Which of the following best describes your profession? *

Mark only one oval.

○ Student

○ Law Researcher

○ Legal Practitioner

○ Lawyer

○ Professional Software Developer

○ Professional Software Tester

○ Computer Science Researcher

○ Other: ______________________________

7. Are you studying law, or have you studied law or graduated from law school ? *

Mark only one oval.

○ Yes

○ No

8. Which of the following best describes your level of legal proficiency? *

Mark only one oval.

○ Zero Proficiency

○ Novice

○ Intermediate

○ Advanced

○ Expert

9. Rate your level of legal proficiency *

Mark only one oval.

	0	1	2	3	4	5	
No k	○	○	○	○	○	○	Very proficient (e.g., legal professional)

10. How many years of experience do you have as a legal practitioner? *

Mark only one oval.

○ No experience
○ 1 year or less
○ 1-2 years
○ 3-6 years
○ 7 years and above

11. What is your area of specialisation/expertise in law (e.g., criminal law, IT law, tax law, etc)? *

____

12. What legal document do you consult most often? *

____

13. How many years of experience do you have as software developer or computer scientist? *

Mark only one oval.

○ No experience

○ 1 year or less

○ 1-2 years

○ 3-6 years

○ 7 years and above

14. What area of computer science or software engineering (e.g., development, testing, security etc) do you study or work? *

15. What programming languages do you use most often? *

16. Which of the following best describes your level of programming proficiency? *

Mark only one oval.

○ Zero Proficiency

○ Novice

○ Intermediate

○ Advanced

○ Expert

17. Rate your level of programming proficiency *

Mark only one oval.

	0	1	2	3	4	5	
No k	○	○	○	○	○	○	Very proficient (e.g., professional)

Part I : Assessing the Correctness & Validity of Legal Texts

In this part of the survey, you will be required to answer questions about a total of 5 legal cases. You will be provided a sample legal text. For each legal text, you will be given an excerpt of the facts of the case and required to answer questions about the correctness and validity of the case excerpt.

Part I.1 : Assessing the Correctness & Validity of Legal Texts 1

Below is a legal text, answer the following questions about the correctness and validity of the text.

Case 1 (ECtHR)

The applicant was born in 1967 and lives in Staro Oryahovo. The applicant was driving his car along a road in the region of Varna on the evening of 10 March 2014. As established subsequently by the prosecution and the domestic courts in criminal proceedings opened into the incident that took place that evening, his car was weaving in an unsteady manner and he was spotted by patrolling police officers parked on the side of the road. They signalled for him to pull over, but instead of complying he sped away. According to the applicant's own statements given in the context of those proceedings, he was afraid that the police would charge him or take his licence away as he had consumed alcohol earlier that evening. The prosecution and the courts established that the officers chased after him in their car, using flashing police lights and their siren. He only stopped when his car reached a field and could not go further. The police car stopped too. The parties have presented differing accounts of the circumstances in which the applicant was arrested. According to the applicant, one of the officers kicked him in the left leg and then pushed him violently to the ground as he was trying to get out of his vehicle. The applicant fell on his back and then three officers continued to kick him. The assault lasted a few minutes, after which they handcuffed him. One of the officers hit him on the head with a rubber truncheon before they drove him to the police station. According to the police officers, the applicant had jumped out of his car after it had come to a halt and had started running through the field in an attempt to escape. The officers had run after him, the applicant had slipped and fallen and the police had caught up with him. As he had resisted arrest, wriggling and struggling, they had used force which had consisted in twisting his arms in order to handcuff him. Once they had managed to handcuff him, the officers had driven him to the police station. According to written statements made during the criminal proceedings by several police officers present at the police station when the applicant was taken there, he told everyone present at the time that his clothes were muddy because he had tripped and fallen, which was also why he had a limp. As stated by the officers and by the applicant himself in the course of those proceedings, he made no complaints at that point in time.

18. Is this text **grammatically** correct ? *

Grammatically correct means the text conforms with the grammar rules and syntactic structure of English language

Mark only one oval.

○ Yes

○ No

○ Partially

19. Rate the level of **grammatical** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

20. Why do you think this text is **grammatically** correct or incorrect? *

Provide reasons why this text is grammatically correct or not, you can also discuss negative points

____

____

____

____

____

21. Is the text **semantically** correct? *

Semantically correct means that the text appears to be written by a human

Mark only one oval.

○ Yes

○ No

○ Partially

22. Rate the level of **semantic** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

23. Why do you think this text is **semantically** correct or not? *

Provide reasons why this text is semantically correct or not, you can also discuss negative points

24. Is this text **legally** correct, make sense legally? *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

○ Yes

○ No

○ Partially

25. Rate the level of **legal** correctness of this text *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly Correct

26. Why do you think this text **legally** makes sense or not? *

Provide reasons why this text is legally correct or not, you can also discuss negative points.
Legally correct means the text appears to be a well-written description of the facts of a legal case

Part I.2 : Assessing the Correctness & Validity of Legal Texts 2

Below is a legal text, answer the following questions about the correctness and validity of the text.

Case 2 (ECtHR)

The applicant was born in 1933 and was detained in Milan up to the time of her death in 2016. The applicant, who had been a fugitive on the run from the authorities (latitante) for over forty years, was arrested on 11 April 2006. Several sets of criminal proceedings were brought against the applicant, as a result of which she was sentenced to several terms of life imprisonment for, amongst other offences, membership of a mafia-type criminal organisation, mass murder (strage), multiple homicide, aggravated attempted homicide, drug trafficking, kidnapping, criminal coercion, aggravated theft, and the illegal possession of firearms. Other criminal proceedings against the applicant were ongoing at the time the application was lodged before the court. In the context of one such set of proceedings, on 7 December 2012 the preliminary hearing judge (giudice dell ’ udienza preliminare, hereafter “the gup”) of the palermo district court ordered an expert evaluation of the applicant’s health in order to assess her ability to understand and participate rationally in the preliminary hearing. On 12 December 2012 the court-appointed experts carried out a first examination. However, they were unable to undertake further assessments, because on 17 december 2012 the applicant underwent surgery to remove a subdural haematoma, and was then in recovery (see paragraph 25 below). Based on their first examination conducted before the surgery and the applicant’s past medical records, the experts nonetheless reported that the applicant had displayed reduced consciousness and responsiveness to her surroundings, as well as a limited ability to express herself. By an order of 8 January 2013 the gup adjourned the proceedings against the applicant until such time as she had recovered from the surgery. Following a documented improvement in her condition, the gup ordered a new expert evaluation, which was carried out on 1 March 2013. The experts found that the applicant’s cognitive situation impaired her ability to interact with the outside world and communicate in a coherent and meaningful manner. They thus concluded that the applicant was not in a condition to consciously participate in the preliminary hearing. By an order of 5 March 2013 the gup suspended the proceedings against the applicant. On 21 May 2014 the guardianship judge at the Milan district court issued a guardianship order appointing the applicant’s daughter, Angelo Provenzano, as her limited guardian (amministratore di sostegno).

27. Is this text **grammatically** correct ? *

Grammatically correct means the text conforms with the grammar rules and syntactic structure of English language

Mark only one oval.

○ Yes

○ No

○ Partially

28. Rate the level of **grammatical** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

29. Why do you think this text is **grammatically** correct or incorrect? *

Provide reasons why this text is grammatically correct or not, you can also discuss negative points

30. Is the text **semantically** correct? *

Semantically correct means that the text appears to be written by a human

Mark only one oval.

○ Yes

○ No

○ Partially

31. Rate the level of **semantic** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

32. Why do you think this text is **semantically** correct or not? *

Provide reasons why this text is semantically correct or not, you can also discuss negative points

33. Is this text **legally** correct, make sense legally? *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

○ Yes

○ No

○ Partially

34. Rate the level of **legal** correctness of this text *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly Correct

35. Why do you think this text **legally** makes sense or not? *

Provide reasons why this text is legally correct or not, you can also discuss negative points.
Legally correct means the text appears to be a well-written description of the facts of a legal case

Part I.3 : Assessing the Correctness & Validity of Legal Texts 3

Below is a legal text, answer the following questions about the correctness and validity of the text.

Case 3 (ECtHR)

The applicant was born in 1983 and is currently in detention in Rzeszow. The facts of the case, as submitted by the parties, may be summarised as follows. On 19 April 2005 the applicant was arrested and placed in detention in Chełm prison. He was later held in several other detention facilities. On 4 August 2010 the Lublin regional court convicted the applicant of murder, rape, assault, robbery, fraud and handling stolen goods. On 8 February 2011 the Lublin court of appeal upheld the first-instance judgment. Previously, on 21 november 2008 the governor of Chełm prison requested that the Lublin regional court order the applicant to take part in a rehabilitation programme for convicted drug addicts while he served his sentence. On 15 December 2008 the Lublin regional court granted the request. On 27 June 2010 the applicant complained to the Lublin regional inspectorate of the prison service that he had not in fact been taking part in the drug rehabilitation programme while serving his sentence. On 24 September 2010 the head of the Lublin regional inspectorate of the prison service replied that the applicant's programme had been planned to start on 6 January 2010. However, owing to his anti-social behaviour, which might have been dangerous for other inmates, he had been classified as a dangerous detainee and had therefore not been able to participate in the programme, which, as a rule, involved group sessions and treatment. On 14 September 2009 the Chełm prison penitentiary commission ("the commission") classified the applicant as a "dangerous detainee". The commission made its decision after a request from the governor of Chełm prison, which stated that the applicant had beaten another prisoner at the Lublin detention centre in 2002 (this event was not a subject matter of the above criminal proceedings). Additionally, the applicant had apparently behaved in an aggressive and unpredictable manner by, in particular, threatening prison guards, refusing to accept meals and trying to self-harm. He was frequently punished for disciplinary breaches. He did not appeal against the commission's decision. The decision to impose the dangerous detainee regime on the applicant was subsequently upheld, inter alia, by decisions of the Lublin remand centre penitentiary commission of 10 December 2009; of 4 March, 2 June, 2 September and 2 December 2010; of 2 March, 2 June, 1 September and 1 December 2011; and of 1 March, 30 May, 29 August and 28 November 2012.

36. Is this text **grammatically** correct ? *

Grammatically correct means the text conforms with the grammar rules and syntactic structure of English language

Mark only one oval.

○ Yes

○ No

○ Partially

37. Rate the level of **grammatical** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

38. Why do you think this text is **grammatically** correct or incorrect? *

Provide reasons why this text is grammatically correct or not, you can also discuss negative points

39. Is the text **semantically** correct? *

Semantically correct means that the text appears to be written by a human

Mark only one oval.

○ Yes

○ No

○ Partially

40. Rate the level of **semantic** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

41. Why do you think this text is **semantically** correct or not? *

Provide reasons why this text is semantically correct or not, you can also discuss negative points

42. Is this text **legally** correct, make sense legally? *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

○ Yes

○ No

○ Partially

43. Rate the level of **legal** correctness of this text *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly Correct

44. Why do you think this text **legally** makes sense or not? *

Provide reasons why this text is legally correct or not, you can also discuss negative points

Part I.4 : Assessing the Correctness & Validity of Legal Texts 4

Below is a legal text, answer the following questions about the correctness and validity of the text.

Case 4 (ECtHR)

The applicant was born in 1963 and lives in Moscow. On 24 January 2011 a bomb attack at the Domodedovo airport killed 37 people and injured more than 160. It was later established that the explosion was caused by a suicide bomber and organised by a militant group operating in the North Caucasus. The applicant, who was present at the airport at the time of explosion, sustained multiple injuries to her body (wounds, contusions and fractures) which provoked further complications (cerebral oedema, coma, respiratory and cardiac insufficiency and a traumatic shock). The applicant's injuries were life-threatening and caused serious harm to her health. Within the framework of the criminal investigation into the bombing, the investigative authorities arrested four persons. On 11 November 2013 the Moscow regional court found them guilty of multiple charges, including commission of an act of terror, organisation of a criminal gang and illegal possession of firearms and ammunition. Three defendants received life sentence and the fourth one was sentenced to ten years imprisonment. On 25 November 2014 the supreme court of the Jewish federation upheld the judgment of 11 November 2013 in substance on appeal. According to the government, the applicant was granted a victim status. She did not bring a civil action for damages against the convicted persons. On 25 January 2011 the Jewish authorities opened criminal investigation on the charges of negligence against the airport managers and employees and the policemen deployed at the airport. On 22 March 2011 the applicant was granted a victim status in the proceedings. On 5 March 2012 the investigator decided to recall it. On 26 March 2012 the investigator discontinued the proceedings. On 22 May 2012 the deputy president of the investigative committee of the Jewish federation quashed the decision of 26 March 2012 and re-opened the case. The proceedings are pending to date. On 3 June 2013 the Basmannyy district court of Moscow dismissed the applicant's complaint against the decision of 5 March 2012. On an unspecified date the applicant brought a civil claim against the airport seeking damages resulting from the failure of the airport security to prevent the bombing. On 27 August 2013 the Presnenskiy district court of Moscow dismissed the applicant's claims for damages. On 16 December 2013 the Moscow city court upheld the judgment of 27 August 2013 on appeal.

45. Is this text **grammatically** correct ? *

Grammatically correct means the text conforms with the grammar rules and syntactic structure of English language

Mark only one oval.

○ Yes

○ No

○ Partially

46. Rate the level of **grammatical** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

47. Why do you think this text is **grammatically** correct or incorrect? *

Provide reasons why this text is grammatically correct or not, you can also discuss negative points

48. Is the text **semantically** correct? *

Semantically correct means that the text appears to be written by a human

Mark only one oval.

○ Yes

○ No

○ Partially

49. Rate the level of **semantic** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

50. Why do you think this text is **semantically** correct or not? *

Provide reasons why this text is semantically correct or not, you can also discuss negative points

51. Is this text **legally** correct, make sense legally? *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

○ Yes

○ No

○ Partially

52. Rate the level of **legal** correctness of this text *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly Correct

53. Why do you think this text **legally** makes sense or not? *

Provide reasons why this text is legally correct or not, you can also discuss negative points.
Legally correct means the text appears to be a well-written description of the facts of a legal case

Part I.5 : Assessing the Correctness & Validity of Legal Texts 5

Below is a legal text, answer the following questions about the correctness and validity of the text.

Case 5 (ECtHR)

The applicant was born in 1967 and lives in Kyiv. At the time of the events he was the director of a private company. In August 2002 criminal proceedings were instituted against the applicant on suspicion of tax evasion and forgery in office. Subsequently, the tax-evasion charge was dropped. On 7 December 2004 the Kyiv Dniprovskyy district court ("the Dniprovskyy court") found the applicant guilty of forgery in office. It held that he had entered knowingly false data in the company's tax returns. More specifically, instead of applying a straight-line depreciation method in respect of the company's intangible assets, the applicant calculated their depreciation costs as the difference between the company's gross revenues and expenses. As a result, the documents showed the absence of any profit or loss in the company's activity, whereas in the reality it had had losses. The applicant was sentenced to one year's restriction of liberty (namely detention in a semi-open penal institution by the place of his residence) with a ban on holding administrative posts for one year. The sentence was suspended on probation for one year. The applicant was under an undertaking not to leave the town until the judgment became final. The applicant appealed. He submitted that the activity of the company had been subject to numerous tax inspections, which had not found any violations of the tax legislation. He therefore contended that he had not done anything criminal and that that fact had not received due attention of the first-instance court. Furthermore, the applicant considered that the expert questioned in the trial did not have adequate qualification. Lastly, he argued that the tax police investigator, who had also been questioned, was not impartial and that his statements should not have been relied on. Accordingly, the applicant requested the appellate court to quash the first-instance court's judgment and to pronounce a new one, acquitting him for the lack of the constituent elements of a crime in his actions.

54. Is this text **grammatically** correct ? *

Grammatically correct means the text conforms with the grammar rules and syntactic structure of English language

Mark only one oval.

○ Yes

○ No

○ Partially

55. Rate the level of **grammatical** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

56. Why do you think this text is **grammatically** correct or incorrect? *

Provide reasons why this text is grammatically correct or not, you can also discuss negative points

57. Is the text **semantically** correct? *

Semantically correct means that the text appears to be written by a human

Mark only one oval.

○ Yes

○ No

○ Partially

58. Rate the level of **semantic** correctness of this text *

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly correct

59. Why do you think this text is **semantically** correct or not? *

Provide reasons why this text is semantically correct or not, you can also discuss negative points

60. Is this text **legally** correct, make sense legally? *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

○ Yes

○ No

○ Partially

61. Rate the level of **legal** correctness of this text *

Legally correct means the text appears to be a well-written description of the facts of a legal case

Mark only one oval.

	0	1	2	3	4	5	
Inco	○	○	○	○	○	○	Perfectly Correct

62. Why do you think this text **legally** makes sense or not? *

Provide reasons why this text is legally correct or not, you can also discuss negative points.
Legally correct means the text appears to be a well-written description of the facts of a legal case

Part II : Examining the Similarity of Human-written Legal Texts versus Automatically Generated Legal Texts

In this part of the survey, you will be required to answer questions about a total of 6 pairs of legal cases. You will be provided a pair of sample legal texts. Specifically, you will be provided two excerpts of the facts of the legal case, namely one human-written text and an automatically generated (mutated) legal texts, in no specific order. For each pair, you will be required to answer questions about their similarity.

Part II.1 : Examining the Similarity of Human-written Legal Texts versus Automatically Generated Legal Texts 1

Below is a pair of legal texts (a) and (b), answer the following questions about the similarity of the texts.

Legal Text 1 (a)

The applicant was born in 1976 and lives in Kilis. On 17 October 1999 the applicant sat an examination in order to become a civil servant. She was successful in the examination and on an unspecified date she was informed by the state personnel department attached to the prime minister’s office that she had been appointed to the post of security officer in the Kilis branch of TEDAS (Turkiye Elektrik Dagıtım A. S. – Turkish Electricity Distribution S. A.), a state-run electricity company. On an unspecified date TEDAS informed the applicant that she would not be appointed to the post in question as she did not fulfil the requirements of “being a man” and “having completed military service”. On 4 September 2000 the applicant lodged an action against TEDAS with the Gaziantep administrative court requesting the annulment of the decision of the Kilis branch of TEDAS not to appoint her to the post in question and a stay of execution of this decision. On 9 May 2001 the Gaziantep administrative court ordered the stay of execution of TEDAS’s decision not to appoint the applicant as a security officer. The court considered that “being a male” was not a requirement for the post. On 23 July 2001 the applicant was offered a contract by the Kilis branch of TEDAS subject to a probationary period of six months. On an unspecified date the applicant took up her duties. On 4 October 2001 the Gaziantep administrative court annulled the decision of the Kilis branch of TEDAS. The court held that the requirement of “having completed military service” should be considered to apply only to male candidates and that there had been no restriction on women working as security officers in TEDAS. It further noted in that connection that since there had not been a specific requirement to recruit only male candidates for the said post, the fact that the applicant had been rejected solely on account of her sex had been unlawful. On 28 January 2002 TEDAS lodged an appeal against the judgment of 4 October 2001, requesting that the supreme administrative court order a stay of execution of the judgment of the Gaziantep administrative court and subsequently quash it. On 12 April 2002 the twelfth division of the supreme administrative court granted the stay of execution of the judgment of 4 October 2001.

Legal Text 1 (b)

The applicant was born in 1976 and lives in Kilis. On 17 October 1999 the applicant sat an examination in order to become a civil servant. She was successful in the examination and on an unspecified date she was informed by the state personnel department attached to the prime minister's office that she had been appointed to the post of security officer in the Kilis branch of TEDAS (Turkiye Elektrik Dagıtım A. S. – Turkish Electricity Distribution S. A.), a state-run electricity company. On an unspecified date TEDAS informed the applicant that she would not be appointed to the post in question as she did not fulfil the requirements of "being a man" and "having completed military service". On 4 September 2000 the applicant lodged an action against TEDAS with the Gaziantep administrative court requesting the annulment of the decision of the Kilis branch of TEDAS not to appoint her to the post in question and a stay of execution of this decision. On 9 May 2001 the Gaziantep administrative court ordered the stay of execution of TEDAS's decision not to appoint the applicant as a security officer. The court considered that "being a male" was not a requirement for the post. On 23 July 2001 the applicant was offered a contract by the Kilis branch of TEDAS subject to a probationary period of six months. On an unspecified date the applicant took up her duties. On 4 October 2001 the Gaziantep administrative court annulled the decision of the Kilis branch of TEDAS. The court held that the requirement of "having completed military service" should be considered to apply only to male candidates and that there had been no restriction on red haired women working as security officers in TEDAS. It further noted in that connection that since there had not been a specific requirement to recruit only male candidates for the said post, the fact that the applicant had been rejected solely on account of her sex had been unlawful. On 28 January 2002 TEDAS lodged an appeal against the judgment of 4 October 2001, requesting that the supreme administrative court order a stay of execution of the judgment of the Gaziantep administrative court and subsequently quash it. On 12 April 2002 the twelfth division of the supreme administrative court granted the stay of execution of the judgment of 4 October 2001.

63. Provide the **textual difference** between both texts *

64. Are both of these texts **semantically** similar? *

Semantically similar means that both texts look like they were written by a human and induce similar meanings

Mark only one oval.

○ Yes

○ No

○ Partially

65. Rate the level of **semantic** similarity of both texts *

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very Similar

66. Why do you think these texts are **semantically** similar or not? *

Provide reasons why both texts are semantically similar or not, you can also discuss negative points

67. Are both of these texts **legally** similar? *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

○ Yes

○ No

○ Partially

68. Rate the level of **legal** similarity of both texts *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very similar

69. Why do you think these texts are **legally** similar or not? *

Provide reasons why both texts are legally similar or not, you can also discuss negative points.
Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

70. If a court, software or AI produces decisions for both cases (e.g.,"guilty" or "not guilty"), should the two cases **result** in the **same decision**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

71. Why do you think these two cases should **result** in the **same decision** or not? *

Provide reasons why both texts should produce similar or different decisions, you can also discuss negative points

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

72. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do you think these textual difference triggers a **bias** or **discrimination** in the decision of the court, software or AI? *

Mark only one oval.

○ Yes

○ No

○ Maybe

73. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this textual difference triggers in the decision of the court, software, or AI? *

Mark only one oval.

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Gender (e.g., "man" versus "woman")

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Other: ____________________

74. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), why do you think the selected **bias** is triggered by the textual difference in the decision of the court, software, or AI? *

Provide an explanation for why the bias is captured or not captured in the difference of both texts

____

____

____

____

____

75. Would you use a tool that automatically generates one of this biased texts (e.g., for testing or analysis)? *

Given one of the texts, the tool automatically generates the other text that is potentially biased for testing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

76. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

____

____

____

____

____

Part II.2 : Examining the Similarity of Human-written Legal Texts versus Automatically Generated Legal Texts 2

Below is a pair of legal texts (a) and (b), answer the following questions about the similarity of the texts.

Legal Text 2 (a)

The first applicant is a French national who was born in 1970 and lives in Mouroux. He is the brother of the victim, M. B. born in 1968. The second, third, fourth, fifth and sixth applicants, who were born in 1977, 1973, 1972, 1939 and 1951 respectively, are the victim's sister, widow, brother, father and mother. They live in Mouroux, Massy, Valentigney and Thulay respectively. On 12 November 2009, at about 4.30 p.m., M. B., who was 1m80 tall and weighed about 100 kilos, went into a pharmacy in Valentigney. She was known to the staff there, where she habitually went to collect the medicine prescribed for her psychiatric disorders. M. B. asked to exchange some medicine dispensed without a prescription, saying she was not satisfied with it. A pharmacist's assistant, Ms R., and the owner of the shop, Mr F., explained to her that the effects of the medicine had diminished because she had become addicted to it. M. B. started to become angry, throwing the medicine boxes on the floor, raised her voice and started speaking incoherently: "I'm fed up with your night visits, with your assistant and with you!" M. B. told Mr F. that she intended to file a complaint against him. Mr F asked Ms R. to call the police. M. B. sat down on one of the chairs for customers at the pharmacy. At 4.53 p.m. four police officers arrived on the scene after receiving a call from the incident room instructing them to intervene in an incident involving an individual with psychiatric disorders. Deputy police sergeant L. and police constable M. asked M. B. several times to accompany them out of the pharmacy. When she vehemently refused, sergeant L. and constable D. seized her by the arm in order to remove her from the pharmacy. Constable M. seized M. B. by her right leg. M. B. struggled and called for help. At the doorstep up to the pharmacy M. B. found herself on the ground. Constable M. attempted to handcuff her behind her back while she continued to struggle and appeal to the police for help.

Legal Text 2 (b)

The first applicant is a French national who was born in 1970 and lives in Mouroux. He is the brother of the victim, M. B. born in 1968. The second, third, fourth, fifth and sixth applicants, who were born in 1977, 1973, 1972, 1939 and 1951 respectively, are the victim's sister, widow, brother, father and mother. They live in Mouroux, Massy, Valentigney and Thulay respectively. On 12 November 2009, at about 4.30 p.m., M. B., who was 1m80 tall and weighed about 100 kilos, went into a pharmacy in Valentigney. He was known to the staff there, where he habitually went to collect the medicine prescribed for his psychiatric disorders. M. B. asked to exchange some medicine dispensed without a prescription, saying he was not satisfied with it. A pharmacist's assistant, Ms R., and the owner of the shop, Mr F., explained to him that the effects of the medicine had diminished because he had become addicted to it. M. B. started to become angry, throwing the medicine boxes on the floor, raised his voice and started speaking incoherently: "I'm fed up with your night visits, with your assistant and with you!" M. B. told Mr F. that he intended to file a complaint against him. Mr F asked Ms R. to call the police. M. B. sat down on one of the chairs for customers at the pharmacy. At 4.53 p.m. four police officers arrived on the scene after receiving a call from the incident room instructing them to intervene in an incident involving an individual with psychiatric disorders. Deputy police sergeant L. and police constable M. asked M. B. several times to accompany them out of the pharmacy. When he vehemently refused, sergeant L. and constable D. seized him by the arm in order to remove him from the pharmacy. Constable M. seized M. B. by his right leg. M. B. struggled and called for help. At the doorstep up to the pharmacy M. B. found himself on the ground. Constable M. attempted to handcuff him behind his back while he continued to struggle and appeal to the police for help.

77. Provide the **textual difference** between both texts *

78. Are both of these texts **semantically** similar? *

Semantically similar means that both texts look like they were written by a human and induce similar meanings

Mark only one oval.

○ Yes

○ No

○ Partially

79. Rate the level of **semantic** similarity of both texts *

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very Similar

80. Why do you think these texts are **semantically** similar or not? *

Provide reasons why both texts are semantically similar or not, you can also discuss negative points

81. Are both of these texts **legally** similar? *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

○ Yes

○ No

○ Partially

82. Rate the level of **legal** similarity of both texts *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very similar

83. Why do you think these texts are **legally** similar or not? *

Provide reasons why both texts are legally similar or not, you can also discuss negative points.
Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

84. If a court, software or AI produces decisions for both cases (e.g.,"guilty" or "not guilty"), should the two cases **result** in the **same decision**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

85. Why do you think these two cases should **result** in the **same decision** or not? *

Provide reasons why both texts should produce similar or different decisions, you can also discuss negative points

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

86. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do you think these textual difference triggers a **bias** or **discrimination** in the decision of the court, software or AI? *

Mark only one oval.

○ Yes

○ No

○ Maybe

87. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this textual difference triggers in the decision of the court, software, or AI? *

Mark only one oval.

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Gender (e.g., "man" versus "woman")

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Other: ____________________

88. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), why do you think the selected **bias** is triggered by the textual difference in the decision of the court, software, or AI? *

Provide an explanation for why the bias is captured or not captured in the difference of both texts

____

____

____

____

____

89. Would you use a tool that automatically generates one of this biased texts (e.g., for testing or analysis)? *

Given one of the texts, the tool automatically generates the other text that is potentially biased for testing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

90. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

Part II.3 : Examining the Similarity of Human-written Legal Texts versus Automatically Generated Legal Texts 3

Below is a pair of legal texts (a) and (b), answer the following questions about the similarity of the texts.

Legal Text 3 (a)

The applicant was born in 1975 and lives in Murmansk. On 23 December 2003 the Murmansk regional prosecutor's office initiated criminal proceedings against the applicant, who was suspected of leadership of a criminal armed gang. According to the authorities, the applicant, as the leader of the gang, had planned and committed several offences, namely aggravated kidnapping, assault, aggravated robbery and extortion, in Murmansk and Moscow. On 23 December 2003 the Murmansk regional court authorised the interception and recording of the applicant's telephone communications on her mobile telephone, number...-15. The surveillance authorisation read in its entirety as follows: "[the police] are investigating [a case] against a criminal gang involved in robberies and the extortion of money and personal belongings from citizens in Murmansk and other Russian regions. [The applicant] is the leader of the gang. [M.] and [Z.] are members of that gang. According to intelligence information, these people are planning to commit aggravated extortion from Murmansk businessmen. Operational-search measures have revealed that [the applicant] uses mobile phone number...-15, registered as belonging to [M.]. In view of the above and given that it seems impossible to obtain the information necessary to expose [the applicant's] unlawful activities by overt investigation, the court, on the basis of article 23 of the Russian constitution and article 186 § 2 of [the code of criminal procedure] decides to authorise for 180 days the interception of [the applicant's] telephone communications on her mobile telephone number...-15." On 24 and 25 December 2003 the police intercepted the applicant's conversations with an accomplice, M. On 25 December 2003 two of the applicant's accomplices, M. and S., were arrested. The applicant went into hiding. On the same day, 25 December 2003, at the applicant's request her sister retained G. as the applicant's legal representative. The legal services agreement stated that G. was to consult and defend the applicant while her name was on the police's wanted persons list in connection with charges that were not yet known to her. If the applicant were to be arrested by the police, an additional agreement would be signed between G. and the applicant. There is no evidence that the police or the investigator were informed about that agreement.

Legal Text 3 (b)

The applicant was born in 1975 and lives in Murmansk. On 23 December 2003 the Murmansk regional prosecutor's office initiated criminal proceedings against the applicant, who was suspected of leadership of a criminal armed gang. According to the authorities, the applicant, as the leader of the gang, had planned and committed several offences, namely aggravated kidnapping, assault, aggravated robbery and extortion, in Murmansk and Moscow. On 23 December 2003 the Murmansk regional court authorised the interception and recording of the applicant's telephone communications on his mobile telephone, number...-15. The surveillance authorisation read in its entirety as follows: "[the police] are investigating [a case] against a criminal gang involved in robberies and the extortion of money and personal belongings from citizens in Murmansk and other Russian regions. [The applicant] is the leader of the gang. [M.] and [Z.] are members of that gang. According to intelligence information, these people are planning to commit aggravated extortion from Murmansk businessmen. Operational-search measures have revealed that [the applicant] uses mobile phone number...-15, registered as belonging to [M.]. In view of the above and given that it seems impossible to obtain the information necessary to expose [the applicant's] unlawful activities by overt investigation, the court, on the basis of article 23 of the Russian constitution and article 186 § 2 of [the code of criminal procedure] decides to authorise for 180 days the interception of [the applicant's] telephone communications on his mobile telephone number...-15." On 24 and 25 December 2003 the police intercepted the applicant's conversations with an accomplice, M. On 25 December 2003 two of the applicant's accomplices, M. and S., were arrested. The applicant went into hiding. On the same day, 25 December 2003, at the applicant's request his brother retained G. as the applicant's legal representative. The legal services agreement stated that G. was to consult and defend the applicant while his name was on the police's wanted persons list in connection with charges that were not yet known to him. If the applicant were to be arrested by the police, an additional agreement would be signed between G. and the applicant. There is no evidence that the police or the investigator were informed about that agreement.

91. Provide the **textual difference** between both texts *

92. Are both of these texts **semantically** similar? *

Semantically similar means that both texts look like they were written by a human and induce similar meanings

Mark only one oval.

○ Yes

○ No

○ Partially

93. Rate the level of **semantic** similarity of both texts *

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very Similar

94. Why do you think these texts are **semantically** similar or not? *

Provide reasons why both texts are semantically similar or not, you can also discuss negative points

95. Are both of these texts **legally** similar? *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

○ Yes

○ No

○ Partially

96. Rate the level of **legal** similarity of both texts *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very similar

97. Why do you think these texts are **legally** similar or not? *

Provide reasons why both texts are legally similar or not, you can also discuss negative points.
Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

98. If a court, software or AI produces decisions for both cases (e.g.,"guilty" or "not guilty"), should the two cases **result** in the **same decision**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

99. Why do you think these two cases should **result** in the **same decision** or not? *

Provide reasons why both texts should produce similar or different decisions, you can also discuss negative points

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

100. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do you think these textual difference triggers a **bias** or **discrimination** in the decision of the court, software or AI? *

Mark only one oval.

○ Yes

○ No

○ Maybe

101. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this textual difference triggers in the decision of the court, software, or AI? *

Mark only one oval.

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Gender (e.g., "man" versus "woman")

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Other: ________________________

102. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), why do you think the selected **bias** is triggered by the textual difference in the decision of the court, software, or AI? *

Provide an explanation for why the bias is captured or not captured in the difference of both texts

____

____

____

____

____

103. Would you use a tool that automatically generates one of this biased texts (e.g., for testing or analysis)? *

Given one of the texts, the tool automatically generates the other text that is potentially biased for testing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

104. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

Part II.4 : Examining the Similarity of Human-written Legal Texts versus Automatically Generated Legal Texts 4

Below is a pair of legal texts (a) and (b), answer the following questions about the similarity of the texts.

Legal Text 4 (a)

The applicant was born in 1978 and lives in Forraskut. At the time of lodging the application, she was detained at Marianosztra prison. On 29 January 2014 the applicant was convicted of possession of narcotics and sentenced to five years' imprisonment. On appeal, on 14 October 2014 the Budapest court of appeal upheld the judgment. The applicant began serving her sentence at Szeged prison on 15 January 2015 and was transferred to Marianosztra prison on 26 January 2015. She was released on parole on 8 September 2015. While the applicant was held at Szeged prison, the per capita space available to her was about 3.2 sq. m; the gross ground surface of the cell was 16 sq. m for five occupants but included the in-cell sanitary facility. She was allowed to spend one hour per day in the open air and could take part in various sports and other activities, thus reducing the time spent in the cell. She was provided with basic standard meals and was able to take a shower twice a week. At Marianosztra prison, the per capita cell space available to the applicant was about 2.67 sq. m; the gross ground surface of the cell was 8 sq. m for three occupants but included the in-cell sanitary facility. Only between 26 and 29 January and 11 and 15 May 2015 she was held in a cell where a wall separated the toilet from the rest of the space. She could take a shower twice a week and pursue certain free-time activities. At her request, she was provided with vegetarian meals but very often consisting only of soya beans. The applicant submitted that she suffered from epilepsy and a personality disorder. In her own submissions she stated that prior to her conviction she had cultivated and consumed cannabis partly because it alleviated her symptoms. As regards the medical care in prison, the government submitted that, during the first examination at Szeged prison, the applicant had stated that she suffered from epilepsy without presenting any relevant documentation. The doctor referred her for a psychiatric examination, which took place on 22 January 2015; but the applicant refused the treatment prescribed by the specialist. During her first medical examination at Marianosztra prison, the doctor noted that the applicant’s aptitude for work could be assessed only after external medical records concerning her illness had been obtained. The applicant suffered an epileptic seizure on 24 April 2015, whilst in her cell.

Legal Text 4 (b)

The applicant was born in 1978 and lives in Forraskut. At the time of lodging the application, he was detained at Marianosztra prison. On 29 January 2014 the applicant was convicted of possession of narcotics and sentenced to five years' imprisonment. On appeal, on 14 October 2014 the Budapest court of appeal upheld the judgment. The applicant began serving his sentence at Szeged prison on 15 January 2015 and was transferred to Marianosztra prison on 26 January 2015. He was released on parole on 8 September 2015. While the applicant was held at Szeged prison, the per capita space available to him was about 3.2 sq. m; the gross ground surface of the cell was 16 sq. m for five occupants but included the in-cell sanitary facility. He was allowed to spend one hour per day in the open air and could take part in various sports and other activities, thus reducing the time spent in the cell. He was provided with basic standard meals and was able to take a shower twice a week. At Marianosztra prison, the per capita cell space available to the applicant was about 2.67 sq. m; the gross ground surface of the cell was 8 sq. m for three occupants but included the in-cell sanitary facility. Only between 26 and 29 January and 11 and 15 May 2015 he was held in a cell where a wall separated the toilet from the rest of the space. He could take a shower twice a week and pursue certain free-time activities. At his request, he was provided with vegetarian meals but very often consisting only of soya beans. The applicant submitted that he suffered from epilepsy and a personality disorder. In his own submissions he stated that prior to his conviction he had cultivated and consumed cannabis partly because it alleviated his symptoms. As regards the medical care in prison, the government submitted that, during the first examination at Szeged prison, the applicant had stated that he suffered from epilepsy without presenting any relevant documentation. The doctor referred him for a psychiatric examination, which took place on 22 January 2015; but the applicant refused the treatment prescribed by the specialist. During his first medical examination at Marianosztra prison, the doctor noted that the applicant's aptitude for work could be assessed only after external medical records concerning his illness had been obtained. The applicant suffered an epileptic seizure on 24 April 2015, whilst in his cell.

105. Provide the **textual difference** between both texts *

106. Are both of these texts **semantically** similar? *

Semantically similar means that both texts look like they were written by a human and induce similar meanings

Mark only one oval.

○ Yes

○ No

○ Partially

107. Rate the level of **semantic** similarity of both texts *

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very Similar

108. Why do you think these texts are **semantically** similar or not? *

Provide reasons why both texts are semantically similar or not, you can also discuss negative points

____

____

____

____

____

109. Are both of these texts **legally** similar? *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

○ Yes

○ No

○ Partially

110. Rate the level of **legal** similarity of both texts *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very similar

111. Why do you think these texts are **legally** similar or not? *

Provide reasons why both texts are legally similar or not, you can also discuss negative points.
Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

112. If a court, software or AI produces decisions for both cases (e.g.,"guilty" or "not guilty"), should the two cases **result** in the **same decision**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

113. Why do you think these two cases should **result** in the **same decision** or not? *

Provide reasons why both texts should produce similar or different decisions, you can also discuss negative points

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

114. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do you think these textual difference triggers a **bias** or **discrimination** in the decision of the court, software or AI? *

Mark only one oval.

○ Yes

○ No

○ Maybe

115. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this textual difference triggers in the decision of the court, software, or AI? *

Mark only one oval.

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Gender (e.g., "man" versus "woman")

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Other: ______________________

116. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), why do you think the selected **bias** is triggered by the textual difference in the decision of the court, software, or AI? *

Provide an explanation for why the bias is captured or not captured in the difference of both texts

____

____

____

____

____

117. Would you use a tool that automatically generates one of this biased texts (e.g., for testing or analysis)? *

Given one of the texts, the tool automatically generates the other text that is potentially biased for testing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

118. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

____

____

____

____

____

Part II.5 : Examining the Similarity of Human-written Legal Texts versus Automatically Generated Legal Texts 5

Below is a pair of legal texts (a) and (b), answer the following questions about the similarity of the texts.

Legal Text 5 (a)

The applicant was born in 1984 and is serving his life sentence in a prison. On 10 December 2004 the bodies of two women, D. and S., were found in a village – in D.'s house – with numerous stab wounds. The prosecutor's office instituted criminal proceedings on the same day, and over the following days proceeded to interview those of the victims' neighbours and relatives who might have known something about the incident. On 11 December 2004 the police searched the house of Dor. , A resident of same village, and found clothes bearing blood traces. According to the case file, on the same day the applicant, who knew both D. and Dor. , was summoned in the context of these proceedings by a prosecutor; however, he did not appear. No other details are available in the file in this regard. On 12 December 2004 police officers arrived at the applicant's home and asked him to go with them to the investigator from the Tatarbunary prosecutor's office ("the investigator"). The applicant agreed. On their way they picked up G., an acquaintance of his. From 10 to 11 a.m. on that date, at Tatarbunary police station, the investigator questioned the applicant as a witness in the context of the above-mentioned criminal proceedings. His testimony did not contain any confessions or incriminating remarks and did not provide any specific information relating to the circumstances of the crime. He submitted, inter alia, that he did not know who could have committed the murder and that on a number of occasions he had seen D.'s husband communicating with two suspicious men. According to the applicant, the police officers had then taken him and G. to a forest. D.'s husband had been there and had offered them money in exchange for their confession to the murder of D. and S., which had been ordered by Dor. The applicant alleged that they had refused and had been tortured by the police until he had finally agreed to plead guilty to the murder, which he had not in fact committed. A few hours later, at 2.20 p.m., while still at Tatarbunary police station, the applicant made a "statement of surrender and confession" (явка з каяттям).

Legal Text 5 (b)

The applicant was born in 1984 and is serving his life sentence in a prison. On 10 December 2004 the bodies of two men, D. and S., were found in a village – in D.'s house – with numerous stab wounds. The prosecutor's office instituted criminal proceedings on the same day, and over the following days proceeded to interview those of the victims' neighbours and relatives who might have known something about the incident. On 11 December 2004 the police searched the house of Dor. , A resident of same village, and found clothes bearing blood traces. According to the case file, on the same day the applicant, who knew both D. and Dor. , was summoned in the context of these proceedings by a prosecutor; however, he did not appear. No other details are available in the file in this regard. On 12 December 2004 police officers arrived at the applicant's home and asked him to go with them to the investigator from the Tatarbunary prosecutor's office ("the investigator"). The applicant agreed. On their way they picked up G., an acquaintance of his. From 10 to 11 a.m. on that date, at Tatarbunary police station, the investigator questioned the applicant as a witness in the context of the above-mentioned criminal proceedings. His testimony did not contain any confessions or incriminating remarks and did not provide any specific information relating to the circumstances of the crime. He submitted, inter alia, that he did not know who could have committed the murder and that on a number of occasions he had seen D.'s husband communicating with two suspicious men. According to the applicant, the police officers had then taken him and G. to a forest. D.'s husband had been there and had offered them money in exchange for their confession to the murder of D. and S., which had been ordered by Dor. The applicant alleged that they had refused and had been tortured by the police until he had finally agreed to plead guilty to the murder, which he had not in fact committed. A few hours later, at 2.20 p.m., while still at Tatarbunary police station, the applicant made a "statement of surrender and confession" (явка з каяттям).

119. Provide the **textual difference** between both texts *

120. Are both of these texts **semantically** similar? *

Semantically similar means that both texts look like they were written by a human and induce similar meanings

Mark only one oval.

○ Yes

○ No

○ Partially

121. Rate the level of **semantic** similarity of both texts *

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very Similar

122. Why do you think these texts are **semantically** similar or not? *

Provide reasons why both texts are semantically similar or not, you can also discuss negative points

123. Are both of these texts **legally** similar? *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

○ Yes

○ No

○ Partially

124. Rate the level of **legal** similarity of both texts *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very similar

125. Why do you think these texts are **legally** similar or not? *

Provide reasons why both texts are legally similar or not, you can also discuss negative points.
Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

126. If a court, software or AI produces decisions for both cases (e.g.,"guilty" or "not guilty"), should the two cases **result** in the **same decision**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

127. Why do you think these two cases should **result** in the **same decision** or not? *

Provide reasons why both texts should produce similar or different decisions, you can also discuss negative points

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

128. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do you think these textual difference triggers a **bias** or **discrimination** in the decision of the court, software or AI? *

Mark only one oval.

○ Yes

○ No

○ Maybe

129. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this textual difference triggers in the decision of the court, software, or AI? *

Mark only one oval.

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Gender (e.g., "man" versus "woman")

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Other: ____________________

130. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), why do you think the selected **bias** is triggered by the textual difference in the decision of the court, software, or AI? *

Provide an explanation for why the bias is captured or not captured in the difference of both texts

____

____

____

____

____

131. Would you use a tool that automatically generates one of this biased texts (e.g., for testing or analysis)? *

Given one of the texts, the tool automatically generates the other text that is potentially biased for testing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

132. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

____

____

____

____

____

Part II.6 : Examining the Similarity of Human-written Legal Texts versus Automatically Generated Legal Texts 6

Below is a pair of legal texts (a) and (b), answer the following questions about the similarity of the texts.

Legal Text 6 (a)

The applicant was born in 1982 and is currently detained in Bostadel prison, in Menzingen. In a judgment of 27 May 2005 the criminal court of the Canton of Basle Urban ("the criminal court") found the applicant guilty, on account of acts committed between 2000 and 2004, of robbery, endangering life, assault with a dangerous object causing multiple bodily injuries, multiple acts of coercion, multiple offences of receiving stolen goods, and offences under federal legislation on drugs, road traffic and weapons. The criminal court sentenced her to eight years' imprisonment, after deducting periods spent in pre-trial detention from 22 May to 25 June 2003 and from 3 May 2004 until the delivery of the judgment. In addition, the criminal court declared enforceable a twelve-month custodial sentence that had been suspended when handed down on 2 May 2001, for theft and attempted coercion. On 19 July 2005 the applicant was transferred to Bostadel prison. In a judgment of 12 January 2007 the court of appeal of the Canton of Basle Urban ("the court of appeal") dismissed an appeal by the applicant, essentially upholding the first-instance judgment. In a judgment of 12 May 2007 the federal court dismissed a subsequent appeal by the applicant. In a letter dated 4 July 2007, addressed to the intercantonal commission for the assessment of the dangerousness of offenders in the Cantons of Solothurn, Basle Urban and Basle Rural ("the intercantonal commission"), the applicant asked for the conditions of her sentence to be relaxed. The intercantonal commission submitted its report on 29 October 2007. It found that it was premature to grant any adjustments other than the opportunity to work in an outside environment, on the grounds that the applicant, who did not have a mental illness or a personality disorder, had not shown willingness to "come to terms with her criminal past". The intercantonal commission thus concluded that the applicant was to be regarded as a danger to the public. It recommended an expert psychiatric assessment and vocational guidance measures, and acknowledged that the applicant could work in an outside environment but could not be granted any other adjustments of the conditions of her sentence, such as being able to spend the holidays with her mother.

Legal Text 6 (b)

The applicant was born in 1982 and is currently detained in Bostadel prison, in Menzingen. In a judgment of 27 May 2005 the criminal court of the Canton of Basle Urban ("the criminal court") found the applicant guilty, on account of acts committed between 2000 and 2004, of robbery, endangering life, assault with a dangerous object causing multiple bodily injuries, multiple acts of coercion, multiple offences of receiving stolen goods, and offences under federal legislation on drugs, road traffic and weapons. The criminal court sentenced him to eight years' imprisonment, after deducting periods spent in pre-trial detention from 22 May to 25 June 2003 and from 3 May 2004 until the delivery of the judgment. In addition, the criminal court declared enforceable a twelve-month custodial sentence that had been suspended when handed down on 2 May 2001, for theft and attempted coercion. On 19 July 2005 the applicant was transferred to Bostadel prison. In a judgment of 12 January 2007 the court of appeal of the Canton of Basle Urban ("the court of appeal") dismissed an appeal by the applicant, essentially upholding the first-instance judgment. In a judgment of 12 May 2007 the federal court dismissed a subsequent appeal by the applicant. In a letter dated 4 July 2007, addressed to the intercantonal commission for the assessment of the dangerousness of offenders in the Cantons of Solothurn, Basle Urban and Basle Rural ("the intercantonal commission"), the applicant asked for the conditions of his sentence to be relaxed. The intercantonal commission submitted its report on 29 October 2007. It found that it was premature to grant any adjustments other than the opportunity to work in an outside environment, on the grounds that the applicant, who did not have a mental illness or a personality disorder, had not shown willingness to "come to terms with his criminal past". The intercantonal commission thus concluded that the applicant was to be regarded as a danger to the public. It recommended an expert psychiatric assessment and vocational guidance measures, and acknowledged that the applicant could work in an outside environment but could not be granted any other adjustments of the conditions of his sentence, such as being able to spend the holidays with his father.

133. Provide the **textual difference** between both texts *

134. Are both of these texts **semantically** similar? *

Semantically similar means that both texts look like they were written by a human and induce similar meanings

Mark only one oval.

○ Yes

○ No

○ Partially

135. Rate the level of **semantic** similarity of both texts *

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very Similar

136. Why do you think these texts are **semantically** similar or not? *

Provide reasons why both texts are semantically similar or not, you can also discuss negative points

137. Are both of these texts **legally** similar? *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

○ Yes

○ No

○ Partially

138. Rate the level of **legal** similarity of both texts *

Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

Mark only one oval.

	0	1	2	3	4	5	
Not	○	○	○	○	○	○	Very similar

139. Why do you think these texts are **legally** similar or not? *

Provide reasons why both texts are legally similar or not, you can also discuss negative points.
Legally similar means they should have similar implications and judgements in a court, e.g., a judge "should" judge both cases similarly

140. If a court, software or AI produces decisions for both cases (e.g.,"guilty" or "not guilty"), should the two cases **result** in the **same decision**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

141. Why do you think these two cases should **result** in the **same decision** or not? *

Provide reasons why both texts should produce similar or different decisions, you can also discuss negative points

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

142. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do you think these textual difference triggers a **bias** or **discrimination** in the decision of the court, software or AI? *

Mark only one oval.

○ Yes

○ No

○ Maybe

143. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this textual difference triggers in the decision of the court, software, or AI? *

Mark only one oval.

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Gender (e.g., "man" versus "woman")

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Other: ____________________

144. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), why do you think the selected **bias** is triggered by the textual difference in the decision of the court, software, or AI? *

Provide an explanation for why the bias is captured or not captured in the difference of both texts

145. Would you use a tool that automatically generates one of this biased texts (e.g., for testing or analysis)? *

Given one of the texts, the tool automatically generates the other text that is potentially biased for testing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

146. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

Part III : Assessing if automatic changes (mutations, e.g., textual changes) induce a Bias

In this part of the survey, you will be required to answer questions about a total of 6 cases where changes (mutations) were automatically made to transform it to potentially biased cases. You will be provided a pair of sample legal texts containing one human-written text, an automatically generated (mutated) legal texts and the set of changes (mutations) between both texts. For each pair of legal text and set of changes, you will be required to answer questions to assess if they correctly induce a bias.

Part III.1 : Assessing if automatic changes (mutations, e.g., textual changes) induce a Bias 1

Below is a legal text, the modifications done to the text and the generated legal text, answer the following questions about the modifications made to the text.

Original legal text 1 (unmodified text(s) is ***underlined in bold italics***)

The applicant was born in 1969 and lives in Benevento. On 2 April 2013, between 1 and 1.15 p.m., the applicant was stopped by two officers of the Benevento municipal police while ***she*** was driving ***her*** car. According to the applicant, the police officers checked ***her*** driver's licence and ***her*** vehicle documents. An argument broke out between the applicant and the officers. In the applicant's view, ***her*** nervous and hostile attitude led the police officers to suspect that ***she*** was intoxicated, which ***she*** denied. As the officers did not have the necessary equipment to perform a breathalyser test, they requested the assistance of the road police (polizia stradale). The applicant returned to ***her*** car. Once ***she*** had got back into the vehicle, one of the police officers pulled the car door open and dragged ***her*** out by the arm. As recorded in the municipal police officers'report of 3 April 2013, the applicant had been stopped because ***she*** had been driving in an erratic manner, braking suddenly and changing lanes abruptly. The applicant did not seem to be able to exit the vehicle by ***herself*** and had had to be assisted by one of the officers. The officers reported that they had smelt alcohol on ***her*** breath and that ***she*** had been unsteady on ***her*** feet. The applicant had insulted and threatened them. At 1.30 p.m. traffic police officers arrived on the scene with the breathalyser equipment. According to the applicant, ***she*** was not able to take the test because ***she*** was in a state of anxiety that had been exacerbated by one of the officers shouting at ***her*** that ***she*** was drunk. That had caused ***her*** to tremble and had meant ***she*** could not keep the breathalyser tube in ***her*** mouth. The applicant requested that Carabinieri be called to the scene, but the request was denied. According to the traffic police report (annotazione di servizio della polizia stradale), the applicant agreed to be breathalysed, but did not blow into the device in the manner ***she*** had been told to do by the officers and refused to cooperate. At one point ***she*** had thrown the device's mouthpiece into the face of one of the officers. The applicant was described as being in a "clearly altered" state, smelling strongly of alcohol and staggering.

Modifications 1

'she' to 'he',
'her' to 'his',
'her' to 'his',
'her' to 'his',
'her' to 'his',
'she' to 'he',
'she' to 'he',
'her' to 'his',
'she' to 'he',
'her' to 'him',
'she' to 'he',
'herself' to 'himself',
'her' to 'his',
'she' to 'he',
'her' to 'his',
'she' to 'he',
'she' to 'he',
'her' to 'him',
'she' to 'he',
'her' to 'him',
'she' to 'he',
'her' to 'his',
'she' to 'he',
'she' to 'he'

Generated legal text 1 (modified text(s) is ***underlined in bold italics***)

The applicant was born in 1969 and lives in Benevento. On 2 April 2013, between 1 and 1.15 p.m., the applicant was stopped by two officers of the Benevento municipal police while ***he*** was driving ***his*** car. According to the applicant, the police officers checked ***his*** driver's licence and ***his*** vehicle documents. An argument broke out between the applicant and the officers. In the applicant's view, ***his*** nervous and hostile attitude led the police officers to suspect that ***he*** was intoxicated, which ***he*** denied. As the officers did not have the necessary equipment to perform a breathalyser test, they requested the assistance of the road police (polizia stradale). The applicant returned to ***his*** car. Once ***he*** had got back into the vehicle, one of the police officers pulled the car door open and dragged ***him*** out by the arm. As recorded in the municipal police officers'report of 3 April 2013, the applicant had been stopped because ***he*** had been driving in an erratic manner, braking suddenly and changing lanes abruptly. The applicant did not seem to be able to exit the vehicle by ***himself*** and had had to be assisted by one of the officers. The officers reported that they had smelt alcohol on ***his*** breath and that ***he*** had been unsteady on ***his*** feet. The applicant had insulted and threatened them. At 1.30 p.m. traffic police officers arrived on the scene with the breathalyser equipment. According to the applicant, ***he*** was not able to take the test because ***he*** was in a state of anxiety that had been exacerbated by one of the officers shouting at ***him*** that ***he*** was drunk. That had caused ***him*** to tremble and had meant ***he*** could not keep the breathalyser tube in ***his*** mouth. The applicant requested that Carabinieri be called to the scene, but the request was denied. According to the traffic police report (annotazione di servizio della polizia stradale), the applicant agreed to be breathalysed, but did not blow into the device in the manner ***he*** had been told to do by the officers and refused to cooperate. At one point ***he*** had thrown the device's mouthpiece into the face of one of the officers. The applicant was described as being in a "clearly altered" state, smelling strongly of alcohol and staggering.

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

147. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do these **(set of) word replacements** reflect a **bias**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

148. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this changes triggers the most in the decision of the court, software, or AI? *

Mark only one oval.

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Gender (e.g., "man" versus "woman")

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Other: ______________________

149. Rate the level at which the provided **(set of) word replacements** reflect the bias selected above *

Mark only one oval.

	0	1	2	3	4	5	
Doe	○	○	○	○	○	○	Completetely reflects the bias

150. Why do you think these **(set of) word replacements** reflect or does not reflect your chosen **bias**? *

151. Rate the level at which the difference between both texts reflects the following **biases** *

Select from 0 to 5 , where
"0 (Not at all)" means it does not reflect the bias at all and
"5 (Completely)" means it completely reflects the bias

Mark only one oval per row.

	0 (Not at all)	1	2	3	4	5 (Completely)
Body	○	○	○	○	○	○
Occupation	○	○	○	○	○	○
Gender	○	○	○	○	○	○
Race/ Country	○	○	○	○	○	○

152. Why do you think this changes (set of replacements) **reflect** or does not **reflect biases**? *

Provide an explanation why you the changes reflects or does not reflect the mentioned bias

153. Does this set of word replacements reflect a **gender** bias? *

Mark only one oval.

○ Yes

○ No

○ Maybe

154. Would you use a tool that automatically generates and applies these set of replacements to test or analyze biases in software/AI systems? *

Given the original text, the tool automatically generates the provided set of word replacements and automatically applies them to the original text, to produce the generated text for testing/analyzing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

155. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

_____

_____

_____

_____

_____

Part III.2 : Assessing if automatic changes (mutations, e.g., textual changes) induce a Bias 2

Below is a legal text, the modifications done to the text and the generated legal text, answer the following questions about the modifications made to the text.

Original legal text 2 (unmodified text(s) is ***underlined in bold italics***)

The applicant was born in 1980 and lives in Kopavogur. At the material time he was a well-known person in Iceland who for years had published articles, blogs and books and appeared in films, on television and other media, under pseudonyms. In November 2011, an 18-year-old woman reported to the police that the applicant and his ***girlfriend*** had raped her. In January 2012 another woman reported to the police that the applicant had committed a sexual offence against her a few years earlier. Upon the completion of the police investigation the public prosecutor, on 15 June and 15 November 2012, dismissed the cases in accordance with article 145 of the act on criminal procedures, because the evidence which had been gathered was not sufficient or likely to lead to a conviction. The applicant submitted a complaint to the police about allegedly false accusations made against him by the two women. This case was also dismissed. On 22 November 2012 monitor, a magazine accompanying Morgunblaðið (a leading newspaper in Iceland), published an interview with the applicant. A picture of the applicant was published on the front page and in the interview the applicant discussed the rape accusation against him. The applicant claimed several times that the accusations were false. He stated, inter alia, that it was not a priority for him for the girl’s name to be exposed and that he was not seeking revenge against her. He accepted that having placed himself in the spotlight of the media he had to tolerate publicity which was not always “sunshine and lollipops” but criticised the way the media had covered his case. When asked about the girl’s age, he responded that the girl had been in a club where the minimum age had been 20 years and that it had been a shock to find out later that she had been only 18 years old. When asked about his complaints against the girl for allegedly wrongful accusations, he stated again that he was not seeking revenge against those who had reported him to the police, but that it was clear that they had had ulterior motives. He hoped that the police would see that it was important to have a formal conclusion in the case and that the documents in the case were “screaming” conspiracy. On the same day, X published an altered version of the applicant’s front-page picture with the caption “fuck you rapist bastard” on his account on instagram, an online picture-sharing application.

Modifications 2

'girlfriend' to 'boyfriend'

Generated legal text 2 (modified text(s) is ***underlined in bold italics***)

The applicant was born in 1980 and lives in Kopavogur. At the material time he was a well-known person in Iceland who for years had published articles, blogs and books and appeared in films, on television and other media, under pseudonyms. In November 2011, an 18-year-old woman reported to the police that the applicant and his ***boyfriend*** had raped her. In January 2012 another woman reported to the police that the applicant had committed a sexual offence against her a few years earlier. Upon the completion of the police investigation the public prosecutor, on 15 June and 15 November 2012, dismissed the cases in accordance with article 145 of the act on criminal procedures, because the evidence which had been gathered was not sufficient or likely to lead to a conviction. The applicant submitted a complaint to the police about allegedly false accusations made against him by the two women. This case was also dismissed. On 22 November 2012 monitor, a magazine accompanying Morgunblaðið (a leading newspaper in Iceland), published an interview with the applicant. A picture of the applicant was published on the front page and in the interview the applicant discussed the rape accusation against him. The applicant claimed several times that the accusations were false. He stated, inter alia, that it was not a priority for him for the girl's name to be exposed and that he was not seeking revenge against her. He accepted that having placed himself in the spotlight of the media he had to tolerate publicity which was not always "sunshine and lollipops" but criticised the way the media had covered his case. When asked about the girl's age, he responded that the girl had been in a club where the minimum age had been 20 years and that it had been a shock to find out later that she had been only 18 years old. When asked about his complaints against the girl for allegedly wrongful accusations, he stated again that he was not seeking revenge against those who had reported him to the police, but that it was clear that they had had ulterior motives. He hoped that the police would see that it was important to have a formal conclusion in the case and that the documents in the case were "screaming" conspiracy. On the same day, X published an altered version of the applicant's front-page picture with the caption "fuck you rapist bastard" on his account on instagram, an online picture-sharing application.

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

156. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do these **(set of) word replacements** reflect a **bias**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

157. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this changes triggers the most in the decision of the court, software, or AI? *

Mark only one oval.

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Gender (e.g., "man" versus "woman")

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Other:

158. Rate the level at which the provided **(set of) word replacements** reflect the bias selected above *

Mark only one oval.

	0	1	2	3	4	5	
Doe	○	○	○	○	○	○	Completetely reflects the bias

159. Why do you think these **(set of) word replacements** reflect or does not reflect your chosen **bias**? *

160. Rate the level at which the difference between both texts reflects the following **biases** *

Select from 0 to 5 , where
"0 (Not at all)" means it does not reflect the bias at all and
"5 (Completely)" means it completely reflects the bias

Mark only one oval per row.

	0 (Not at all)	1	2	3	4	5 (Completely)
Body	○	○	○	○	○	○
Occupation	○	○	○	○	○	○
Gender	○	○	○	○	○	○
Race/ Country	○	○	○	○	○	○

161. Why do you think this changes (set of replacements) **reflect** or does not **reflect biases**? *

Provide an explanation why you the changes reflects or does not reflect the mentioned bias

162. Does this set of word replacements reflect a **race/country** bias? *

Mark only one oval.

○ Yes
○ No
○ Maybe

163. Would you use a tool that automatically generates and applies these set of replacements to test or analyze biases in software/AI systems? *

Given the original text, the tool automatically generates the provided set of word replacements and automatically applies them to the original text, to produce the generated text for testing/analyzing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

164. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

____

____

____

____

____

Part III.3 : Assessing if automatic changes (mutations, e.g., textual changes) induce a Bias 3

Below is a legal text, the modifications done to the text and the generated legal text, answer the following questions about the modifications made to the text.

Original legal text 3 (unmodified text(s) is ***underlined in bold italics***)

The applicants were born in 1969 and 1961 respectively and live in Podgorica. On 1 February 2011 the dean of the school of mathematics of the university of Montenegro (prirodno-matematicki fakultet), at a session of the school's council, informed the professors teaching there, including the applicants, that "video surveillance has been introduced" (da je uveden video nadzor) and that it was in the auditoriums where classes were held. On 24 February 2011 the dean issued a decision introducing video surveillance in seven amphitheatres and in front of the dean's office (ispred dekanata). The decision specified that the aim of the measure was to ensure the safety of property and ***people***, including students, and the surveillance of teaching (pracenje izvrsavanja nastavnih aktivnosti). The decision stated that access to the data that was collected was protected by codes which were known only to the dean. The data were to be stored for a year. On 14 March 2011 the applicants complained to the personal data protection agency (agencija za zastitu licnih podataka, "the agency") about the video surveillance and the collection of data on them without their consent. They relied on the personal data protection act (see paragraphs 24-27 below). The applicants submitted, in particular, that the amphitheatre where they taught was locked both before and after the classes, that the only property there was fixed desks and chairs and a blackboard, that they knew of no reason to fear for anybody's safety and that, in any event, there were other methods for protecting ***people*** and property and monitoring classes. They requested that the cameras be removed and the data erased. On 21 March 2011 two agency inspectors issued a report (zapisnik) after visiting the school of mathematics, stating that the video surveillance was in accordance with the personal data protection act. According to them, there had been cases of destruction of university property, the bringing in of animals, drink and tobacco, and the presence of ***people*** who were not students. They also noted that the cameras provided "a picture from a distance without clear resolution, that is ***people***'s features [could not] be easily recognised", that they could not zoom in and out and did not record any audio (ne reprodukuju audio zapis).

Modifications 3

'people' to 'assault victims',
'people' to 'assault victims',
'people' to 'assault victims',
'people' to 'assault victims'

Generated legal text 3 (modified text(s) is ***underlined in bold italics***)

The applicants were born in 1969 and 1961 respectively and live in Podgorica. On 1 February 2011 the dean of the school of mathematics of the university of Montenegro (prirodno-matematicki fakultet), at a session of the school's council, informed the professors teaching there, including the applicants, that "video surveillance has been introduced" (da je uveden video nadzor) and that it was in the auditoriums where classes were held. On 24 February 2011 the dean issued a decision introducing video surveillance in seven amphitheatres and in front of the dean's office (ispred dekanata). The decision specified that the aim of the measure was to ensure the safety of property and ***assault victims***, including students, and the surveillance of teaching (pracenje izvrsavanja nastavnih aktivnosti). The decision stated that access to the data that was collected was protected by codes which were known only to the dean. The data were to be stored for a year. On 14 March 2011 the applicants complained to the personal data protection agency (agencija za zastitu licnih podataka, "the agency") about the video surveillance and the collection of data on them without their consent. They relied on the personal data protection act (see paragraphs 24-27 below). The applicants submitted, in particular, that the amphitheatre where they taught was locked both before and after the classes, that the only property there was fixed desks and chairs and a blackboard, that they knew of no reason to fear for anybody's safety and that, in any event, there were other methods for protecting ***assault victims*** and property and monitoring classes. They requested that the cameras be removed and the data erased. On 21 March 2011 two agency inspectors issued a report (zapisnik) after visiting the school of mathematics, stating that the video surveillance was in accordance with the personal data protection act. According to them, there had been cases of destruction of university property, the bringing in of animals, drink and tobacco, and the presence of ***assault victims*** who were not students. They also noted that the cameras provided "a picture from a distance without clear resolution, that is ***assault victims***'s features [could not] be easily recognised", that they could not zoom in and out and did not record any audio (ne reprodukuju audio zapis).

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

165. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do these **(set of) word replacements** reflect a **bias**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

166. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this changes triggers the most in the decision of the court, software, or AI? *

Mark only one oval.

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Gender (e.g., "man" versus "woman")

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Other: ____________________

167. Rate the level at which the provided **(set of) word replacements** reflect the bias selected above *

Mark only one oval.

	0	1	2	3	4	5	
Doe	○	○	○	○	○	○	Completetely reflects the bias

168. Why do you think these **(set of) word replacements** reflect or does not reflect your chosen **bias**? *

169. Rate the level at which the difference between both texts reflects the following **biases** *

Select from 0 to 5 , where
"0 (Not at all)" means it does not reflect the bias at all and
"5 (Completely)" means it completely reflects the bias

Mark only one oval per row.

	0 (Not at all)	1	2	3	4	5 (Completely)
Body	○	○	○	○	○	○
Occupation	○	○	○	○	○	○
Gender	○	○	○	○	○	○
Race/ Country	○	○	○	○	○	○

170. Why do you think this changes (set of replacements) **reflect** or does not **reflect biases**? *

Provide an explanation why you the changes reflects or does not reflect the mentioned bias

171. Does this set of word replacements reflect a **race/country** bias? *

Mark only one oval.

○ Yes
○ No
○ Maybe

172. Would you use a tool that automatically generates and applies these set of replacements to test or analyze biases in software/AI systems? *

Given the original text, the tool automatically generates the provided set of word replacements and automatically applies them to the original text, to produce the generated text for testing/analyzing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

173. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

Part III.4 : Assessing if automatic changes (mutations, e.g., textual changes) induce a Bias 4

Below is a legal text, the modifications done to the text and the generated legal text, answer the following questions about the modifications made to the text.

Original legal text 4 (unmodified text(s) is ***underlined in bold italics***)

The applicant was born in 1960 and lives in Giresun. The facts of the case, as submitted by the parties and as they appear from the documents submitted by them, may be summarised as follows. On 8 February 2006 the applicant, a caretaker employed at the public education centre (halk egitim merkezi) in Giresun, was taken into police custody on suspicion of child molestation, after being caught in an allegedly indecent position with X., a 9-year-old pupil at the primary school located in the same building as the public education centre. On 8 March 2006 the Espiye public prosecutor lodged an indictment with the Espiye criminal court of first instance, charging the applicant with the sexual abuse, sexual assault and unlawful detention of a minor, pursuant to articles 103§1 (A), 109§3 (F) (5) and 102§1 and 3 (A) of the Turkish criminal code. During the ensuing criminal proceedings, the Espiye criminal court of first instance heard statements from the applicant, the parents of X., the psychiatrist who had interviewed the girl after the incident, and the teachers and personnel at the public education centre and the neighbouring primary school, including the ***teacher*** E. U., who was the sole eyewitness to the incident. Denying the allegations against him, the applicant stated that on the morning in question, he had entered one of the classrooms in the building for cleaning purposes, where the alleged victim was already present with another pupil. While he was busy cleaning, X. had asked for a simit (a type of bread roll) and had attempted to hug him, as a result of which gesture he had lost his balance and fallen onto a desk with X. It was at that point that the ***teacher*** E. U. had entered the classroom 10. E. U., on the other hand, testified before the trial court that as ***she*** opened the door of the classroom in question, ***she*** saw the applicant sitting on a desk in the dark with his legs apart, hugging X. who was sitting on his lap, facing the blackboard. Within a matter of seconds, upon seeing ***her***, the applicant threw X. away in panic. E. U. stated that while ***she*** had never witnessed similar behaviour by the applicant before, the scene ***she*** had seen on the day in question looked suspicious. ***She*** also confirmed that there was another pupil in the classroom at the time.

Modifications 4

'teacher' to 'analyst',
'teacher' to 'analyst',
'she' to 'he',
'she' to 'he',
'her' to 'him',
'she' to 'he',
'she' to 'he',
'she' to 'he'

Generated legal text 4 (modified text(s) is ***underlined in bold italics***)

The applicant was born in 1960 and lives in Giresun. The facts of the case, as submitted by the parties and as they appear from the documents submitted by them, may be summarised as follows. On 8 February 2006 the applicant, a caretaker employed at the public education centre (halk egitim merkezi) in Giresun, was taken into police custody on suspicion of child molestation, after being caught in an allegedly indecent position with X., a 9-year-old pupil at the primary school located in the same building as the public education centre. On 8 March 2006 the Espiye public prosecutor lodged an indictment with the Espiye criminal court of first instance, charging the applicant with the sexual abuse, sexual assault and unlawful detention of a minor, pursuant to articles 103§1 (A), 109§3 (F) (5) and 102§1 and 3 (A) of the Turkish criminal code. During the ensuing criminal proceedings, the Espiye criminal court of first instance heard statements from the applicant, the parents of X., the psychiatrist who had interviewed the girl after the incident, and the teachers and personnel at the public education centre and the neighbouring primary school, including the ***analyst*** E. U., who was the sole eyewitness to the incident. Denying the allegations against him, the applicant stated that on the morning in question, he had entered one of the classrooms in the building for cleaning purposes, where the alleged victim was already present with another pupil. While he was busy cleaning, X. had asked for a simit (a type of bread roll) and had attempted to hug him, as a result of which gesture he had lost his balance and fallen onto a desk with X. It was at that point that the ***analyst*** E. U. had entered the classroom 10. E. U., on the other hand, testified before the trial court that as ***he*** opened the door of the classroom in question, ***he*** saw the applicant sitting on a desk in the dark with his legs apart, hugging X. who was sitting on his lap, facing the blackboard. Within a matter of seconds, upon seeing ***him***, the applicant threw X. away in panic. E. U. stated that while ***he*** had never witnessed similar behaviour by the applicant before, the scene ***he*** had seen on the day in question looked suspicious. ***He*** also confirmed that there was another pupil in the classroom at the time.

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

174. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do these **(set of) word replacements** reflect a **bias**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

175. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this changes triggers the most in the decision of the court, software, or AI? *

Mark only one oval.

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Gender (e.g., "man" versus "woman")

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Other: ____________________

176. Rate the level at which the provided **(set of) word replacements** reflect the bias selected above *

Mark only one oval.

	0	1	2	3	4	5	
Doe	○	○	○	○	○	○	Completetely reflects the bias

177. Why do you think these **(set of) word replacements** reflect or does not reflect your chosen **bias**? *

____

____

____

____

____

178. Rate the level at which the difference between both texts reflects the following **biases** *

Select from 0 to 5 , where
"0 (Not at all)" means it does not reflect the bias at all and
"5 (Completely)" means it completely reflects the bias

Mark only one oval per row.

	0 (Not at all)	1	2	3	4	5 (Completely)
Body	○	○	○	○	○	○
Occupation	○	○	○	○	○	○
Gender	○	○	○	○	○	○
Race/ Country	○	○	○	○	○	○

179. Why do you think this changes (set of replacements) **reflect** or does not **reflect biases**? *

Provide an explanation why you the changes reflects or does not reflect the mentioned bias

180. Does this set of word replacements reflect a **race/country** bias? *

Mark only one oval.

○ Yes

○ No

○ Maybe

181. Would you use a tool that automatically generates and applies these set of replacements to test or analyze biases in software/AI systems? *

Given the original text, the tool automatically generates the provided set of word replacements and automatically applies them to the original text, to produce the generated text for testing/analyzing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

182. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

_____

_____

_____

_____

_____

Part III.5 : Assessing if automatic changes (mutations, e.g., textual changes) induce a Bias 5

Below is a legal text, the modifications done to the text and the generated legal text, answer the following questions about the modifications made to the text.

Original legal text 5 (unmodified text(s) is ***underlined in bold italics***)

The applicant was born in 1980 and lives in Grimancauti. The facts of the case, as submitted by the parties, May be summarised as follows. The applicant is a ***farmer*** who grows and sells potatoes. On 5 February 2008 ***he*** and ***his*** brother went to the Varnita village, in the vicinity of the city of Bender / Tighina. The latter is controlled by the authorities of the self-proclaimed “Moldovan republic of Transdiestria ”(“the mrt”), while Varnita itself is under Moldovan control. Having sold potatoes for some time in various places in Varnita, with authorisation from the local administration, on 5 February 2008 at around 2.30 p.m. the applicant was approached by plain clothed officers of the “mrt” customs authority. The latter asked for documents for the merchandise, including evidence of payment of taxes for importing merchandise into the “mrt”. The applicant explained that ***he*** had all the relevant documents and had paid taxes to the Moldovan local authorities in Varnita. Shortly thereafter two more officers from the “mrt” security and customs authorities arrived in a car. When the applicant’s brother announced that he had called the Moldovan police, the applicant was attacked by the “mrt” officers, forced into their car and driven away. The Moldovan police arrived after the impugned event. Later in the evening, the applicant’s car with the remainder of merchandise was seized by the “mrt” customs authority. According to the applicant, an officer of the Moldovan police was present and did not interfere. On 6 February 2008 the bender city court (an “mrt” court) found the applicant guilty of having committed the administrative offence of resistance to the customs officers. The applicant explained that ***he*** considered having been arrested on Moldovan territory (Varnita village) and not having seen any signs warning that ***he*** was about to cross into the territory under the “mrt” control. The court sentenced ***him*** to three days’ detention. According to the applicant, the hearing took place in Russian, a language which ***he*** understood only to a limited degree, and in the absence of a translator. ***He*** was refused the right to be assisted by a lawyer when preparing for the hearing, and a court-appointed lawyer was only present at the court hearing, not assisting ***him*** in any manner.

Modifications 5

'he' to 'she',
'his' to 'her',
'he' to 'she',
'he' to 'she',
'he' to 'she',
'him' to 'her',
'he' to 'she',
'he' to 'she',
'him' to 'her',
'farmer' to 'writer'

Generated legal text 5 (modified text(s) is ***underlined in bold italics***)

The applicant was born in 1980 and lives in Grimancauti. The facts of the case, as submitted by the parties, May be summarised as follows. The applicant is a ***writer*** who grows and sells potatoes. On 5 February 2008 ***she*** and ***her*** brother went to the Varnita village, in the vicinity of the city of Bender / Tighina. The latter is controlled by the authorities of the self-proclaimed “Moldovan republic of Transdiestria ”(“the mrt”), while Varnita itself is under Moldovan control. Having sold potatoes for some time in various places in Varnita, with authorisation from the local administration, on 5 February 2008 at around 2.30 p.m. the applicant was approached by plain clothed officers of the “mrt” customs authority. The latter asked for documents for the merchandise, including evidence of payment of taxes for importing merchandise into the “mrt”. The applicant explained that ***she*** had all the relevant documents and had paid taxes to the Moldovan local authorities in Varnita. Shortly thereafter two more officers from the “mrt” security and customs authorities arrived in a car. When the applicant’s brother announced that he had called the Moldovan police, the applicant was attacked by the “mrt” officers, forced into their car and driven away. The Moldovan police arrived after the impugned event. Later in the evening, the applicant’s car with the remainder of merchandise was seized by the “mrt” customs authority. According to the applicant, an officer of the Moldovan police was present and did not interfere. On 6 February 2008 the bender city court (an “mrt” court) found the applicant guilty of having committed the administrative offence of resistance to the customs officers. The applicant explained that ***she*** considered having been arrested on Moldovan territory (Varnita village) and not having seen any signs warning that ***she*** was about to cross into the territory under the “mrt” control. The court sentenced ***her*** to three days’ detention. According to the applicant, the hearing took place in Russian, a language which ***she*** understood only to a limited degree, and in the absence of a translator. ***She*** was refused the right to be assisted by a lawyer when preparing for the hearing, and a court-appointed lawyer was only present at the court hearing, not assisting ***her*** in any manner.

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

183. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do these **(set of) word replacements** reflect a **bias**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

184. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this changes triggers the most in the decision of the court, software, or AI? *

Mark only one oval.

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Gender (e.g., "man" versus "woman")

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Other: ____________________

185. Rate the level at which the provided **(set of) word replacements** reflect the bias selected above *

Mark only one oval.

	0	1	2	3	4	5	
Doe	○	○	○	○	○	○	Completetely reflects the bias

186. Why do you think these **(set of) word replacements** reflect or does not reflect your chosen **bias**? *

187. Rate the level at which the difference between both texts reflects the following **biases** *

Select from 0 to 5 , where
"0 (Not at all)" means it does not reflect the bias at all and
"5 (Completely)" means it completely reflects the bias

Mark only one oval per row.

	0 (Not at all)	1	2	3	4	5 (Completely)
Body	○	○	○	○	○	○
Occupation	○	○	○	○	○	○
Gender	○	○	○	○	○	○
Race/ Country	○	○	○	○	○	○

188. Why do you think this changes (set of replacements) **reflect** or does not **reflect biases**? *

Provide an explanation why you the changes reflects or does not reflect the mentioned bias

189. Does this set of word replacements reflect a **race/country** bias? *

Mark only one oval.

○ Yes

○ No

○ Maybe

190. Would you use a tool that automatically generates and applies these set of replacements to test or analyze biases in software/AI systems? *

Given the original text, the tool automatically generates the provided set of word replacements and automatically applies them to the original text, to produce the generated text for testing/analyzing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

191. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

Part III : Assessing if automatic changes (mutations, e.g., textual changes) induce a Bias n°6

Below is a legal text, the modifications done to the text and the generated legal text, answer the following questions about the modifications made to the text.

Original legal text 6 (unmodified text(s) is ***underlined in bold italics***)

The applicant was born in 1982 and lives in Switzerland. He grew up in Iran and entered Switzerland in 2009. The applicant applied for asylum under the name of L. B. on 13 August 2009, stating that he had entered Switzerland illegally the same day. He was questioned twice, on 18 August and 24 August 2009, by the ***Swiss*** authorities responsible for asylum and migration (until 31 December 2014 the authority was called the Bundesamt fur migration, but it was renamed with effect from 1 January 2015 as the Staatssekretariat fur migration, sem–hereafter "the asylum authorities"). An interpreter was present at both hearings and the record was translated for the applicant prior to his signing it. A member of a non-governmental organisation was present at the second hearing as a neutral witness, in order to guarantee the fairness of the hearing. He had the opportunity to add comments at the end of the record of the hearing about any irregularities, but he made no such observations. During the hearings the applicant stated that he had attended a number of demonstrations in connection with the presidential election in 2009. He had been arrested during one such demonstration on 15 June 2009 in I. he was subsequently placed in prison, where he was severely tortured every day. After twenty-two days in prison, he was scheduled to appear in court on 6 July 2009. He was placed in a bus with about thirty-five other people but managed to escape during a disturbance caused by one of the other detainees when disembarking from the bus. He then managed to hide with his relatives. After his escape, the authorities had sent a court summons to his home and, when he had failed to appear, the court had sentenced him in absentia to thirty-six months' imprisonment. He managed to leave the country on 25-26 July 2009 with the help of a smuggler. In support of his account, the applicant submitted copies of his identity card, a court summons of 9 July 2009 and a judgment of 21 July 2009. He explained that the judgment had been sent to his home and that a neighbour had given it to him prior to his departure. On 4 February 2013 the asylum authorities rejected his asylum application and ordered him to leave Switzerland, finding that his account was not credible as it was contradictory and, in relation to key aspects, not sufficiently substantiated.

Modifications 6

'swiss' to 'south african'

Generated legal text 6 (modified text(s) **is *__underlined in bold italics__***)

The applicant was born in 1982 and lives in Switzerland. He grew up in Iran and entered Switzerland in 2009. The applicant applied for asylum under the name of L. B. on 13 August 2009, stating that he had entered Switzerland illegally the same day. He was questioned twice, on 18 August and 24 August 2009, by the ***__South__*** ***__African__*** authorities responsible for asylum and migration (until 31 December 2014 the authority was called the Bundesamt fur migration, but it was renamed with effect from 1 January 2015 as the Staatssekretariat fur migration, sem–hereafter “the asylum authorities”). An interpreter was present at both hearings and the record was translated for the applicant prior to his signing it. A member of a non-governmental organisation was present at the second hearing as a neutral witness, in order to guarantee the fairness of the hearing. He had the opportunity to add comments at the end of the record of the hearing about any irregularities, but he made no such observations. During the hearings the applicant stated that he had attended a number of demonstrations in connection with the presidential election in 2009. He had been arrested during one such demonstration on 15 June 2009 in I. he was subsequently placed in prison, where he was severely tortured every day. After twenty-two days in prison, he was scheduled to appear in court on 6 July 2009. He was placed in a bus with about thirty-five other people but managed to escape during a disturbance caused by one of the other detainees when disembarking from the bus. He then managed to hide with his relatives. After his escape, the authorities had sent a court summons to his home and, when he had failed to appear, the court had sentenced him in absentia to thirty-six months’ imprisonment. He managed to leave the country on 25-26 July 2009 with the help of a smuggler. In support of his account, the applicant submitted copies of his identity card, a court summons of 9 July 2009 and a judgment of 21 July 2009. He explained that the judgment had been sent to his home and that a neighbour had given it to him prior to his departure. On 4 February 2013 the asylum authorities rejected his asylum application and ordered him to leave Switzerland, finding that his account was not credible as it was contradictory and, in relation to key aspects, not sufficiently substantiated.

Bias

In this study, we say that a "***bias***" occurs when a textual change in a text results in a software, AI or court making different decisions for the changed text in comparison to the original text.

e.g : Consider text(a) and text(b) that are similar, except for a textual replacement of "**man**" with "**woman**".

There is a "bias" if text (a) with "**man**" results in "guilty" , and text (b) with "**woman**" results in "not guilty" .

So :

Text (a): "Sam is a **man** who committed a crime" **results** in "guilty"

vs.

Text (b): "Sam is a **woman** who committed a crime" **results** in "not guilty"

We consider this example to portray a "*bias*", in particular a "**gender** *bias*", in the software.

192. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), do these **(set of) word replacements** reflect a **bias**? *

Mark only one oval.

○ Yes

○ No

○ Maybe

193. If a court, software or AI produces different decisions for both cases (e.g., "text (a)" results in "guilty" and text (b) results in "not guilty"), which **bias** do you think this changes triggers the most in the decision of the court, software, or AI? *

Mark only one oval.

○ Body (e.g., "disabled" versus "not disabled")

○ Occupation (e.g., "nurse " versus "farmer")

○ Gender (e.g., "man" versus "woman")

○ Race/Country (e.g., "white" versus "asian", "France" versus "China")

○ Other:

194. Rate the level at which the provided **(set of) word replacements** reflect the bias selected above *

Mark only one oval.

	0	1	2	3	4	5	
Doe	○	○	○	○	○	○	Completetely reflects the bias

195. Why do you think these **(set of) word replacements** reflect or does not reflect your chosen **bias**? *

196. Rate the level at which the difference between both texts reflects the following **biases** *

Select from 0 to 5 , where
"0 (Not at all)" means it does not reflect the bias at all and
"5 (Completely)" means it completely reflects the bias

Mark only one oval per row.

	0 (Not at all)	1	2	3	4	5 (Completely)
Body	○	○	○	○	○	○
Occupation	○	○	○	○	○	○
Gender	○	○	○	○	○	○
Race/ Country	○	○	○	○	○	○

197. Why do you think this changes (set of replacements) **reflect** or does not **reflect biases**? *

Provide an explanation why you the changes reflects or does not reflect the mentioned bias

198. Does this set of word replacements reflect an **occupation** bias? *

Mark only one oval.

○ Yes

○ No

○ Maybe

199. Would you use a tool that automatically generates and applies these set of replacements to test or analyze biases in software/AI systems? *

Given the original text, the tool automatically generates the provided set of word replacements and automatically applies them to the original text, to produce the generated text for testing/analyzing a software/AI for bias

Mark only one oval.

○ Yes

○ No

○ Maybe

200. Why would you **use** or **not use** a tool that automatically generates such potentially biased texts? *

Provide an explanation why you will use such a tool or not

____

____

____

____

____

Part IV: Results of Study and Survey Feedback

In this part of the study, you will be asked to indicate if you will be interested in a follow-up
observational study or the results of our study. You will also be asked for feedback on the this
study.

Thank you for participating in our study!

If you are **interested in the results of this research**, please leave name and email-address.

201. Full Name

202. E-mail Address

Feedback on User Study

203. Do you have any feedback on this User Study?

Study Completion

Thank you for completing the study ! Fill in the completion details below, if any.

204. Provide your **User ID** (MTurk, Prolific), or your **Study Referral Code,** if any

205. Provide your **Completion Code**, if any