Ethics and Information Technology (2025) 27:16
https://doi.org/10.1007/s10676-025-09821-w

ORIGINAL PAPER

Responsible guidelines for authorship attribution tasks in NLP

Vageesh Saxena[1] · **Aurelia Tamò-Larrieux**[1,2] · **Gijs Van Dijck**[1] · **Gerasimos Spanakis**[1]

Published online: 22 March 2025

Abstract
Authorship Attribution (AA) approaches in Natural Language Processing (NLP) are important in various domains, including forensic analysis and cybercrime. However, they pose Ethical, Legal, and Societal Implications/Aspects (ELSI/ELSA) challenges that remain underexplored. Inspired by foundational AI ethics guidelines and frameworks, this research introduces a comprehensive framework of responsible guidelines that focuses on AA tasks in NLP, which are tailored to different stakeholders and development phases. These guidelines are structured around four core principles: privacy and data protection, fairness and non-discrimination, transparency and explainability, and societal impact. Furthermore, to illustrate a practical application of our guidelines, we apply them to a recent AA study that targets identifying and linking potential human trafficking vendors. We believe the proposed guidelines can assist researchers and practitioners in justifying their decisions, assisting ethical committees in promoting responsible practices, and identifying ethical concerns related to NLP-based AA approaches. Our study aims to contribute to ensuring the responsible development and deployment of AA tools.

Keywords Responsible AI · Authorship attribution (AA) · Natural language processing (NLP) · Privacy & data protection · Fairness & non-discrimination · Transparency & Explainability · Societal impact

Introduction

Authorship Attribution (AA) in Natural Language Processing (NLP) involves examining stylometric and linguistic features from textual segments of multiple authors by analyzing word choice, syntactic patterns, and writing styles. Recent NLP advancements have enabled Machine Learning (ML) approaches for authorship verification Koppel and Schler (2004); Bevendorff et al. (2022); Lei et al. (2022) and identification (Mohsen et al., 2016; Kale & Prasad, 2017; Yuluce & Dalkc, 2022) tasks. These techniques find utility in various domains, including forensic linguistics (Iqbal et al., 2008; Yang & Chow, 2014a; Wright & May, 2014; Fobbe, 2021), cybersecurity (Nirkhi & Dr.R.V.Dharaskar, 2013; Mateless et al., 2021), cybercrime detection (Zhang et al., 2019; Zheng et al., 2003; Kumar et al., 2020; Manolache et al., 2022; Saxena et al., 2023b, a), anomaly detection (Guthrie et al., 2007; Neme et al., 2011; Boukhaled & Ganascia, 2014), and many others.

AA techniques have evolved from traditional stylometric methods to sophisticated ML and deep learning (DL) models, leveraging data to identify unique linguistic fingerprints. Stylometric approaches rely on lexical, syntactic, and character-level features such as word frequency, grammar patterns, and n-grams, offering effective solutions for small datasets requiring high interpretability (Prasad et al., 2015). As textual data complexity grows, ML approaches like Support Vector Machines, Naive Bayes, and Random Forests automate feature learning, enabling scalability for larger datasets (Diederich et al., 2003; Altheneyan & Menai, 2014; Khonji et al., 2015). These methods combine linguistic markers for optimal performance, but their success hinges on quality training data and thoughtful feature engineering (Koppel et al., 2002; Stamatatos, 2009; Sapkota et al., 2015).

The advent of DL models has further advanced AA, with architectures like RNNs, LSTMs (Jafariakinabad

✉ Vageesh Saxena
v.saxena@maastrichtuniversity.nl

Aurelia Tamò-Larrieux
aurelia.tamo-larrieux@unil.ch

Gijs Van Dijck
gijs.vandijck@maastrichtuniversity.nl

Gerasimos Spanakis
jerry.spanakis@maastrichtuniversity.nl

[1] Law & Tech Lab, Maastricht University, Maastricht, Netherlands

[2] Université de Lausanne, Lausanne, Switzerland

et al., 2019), CNNs (Zhang et al., 2015), and transformers such as BERT (Devlin et al., 2019) and GPT (Yenduri et al., 2023). These models capture hierarchical and nuanced text patterns, excelling in large-scale tasks while reducing reliance on manual feature extraction (Radford et al., 2019; Saxena et al., 2023a). However, DL models often lack transparency, posing challenges in domains like forensic linguistics, where interpretability is critical (Rudin, 2019). In contrast, traditional and ML methods remain relevant in cases demanding lower computational resources or high explainability (Strubell et al., 2019). Hybrid approaches combine traditional and modern methods, integrating diverse features into ensemble models to enhance robustness across varied contexts (Salur & Aydin, 2020; Stamatatos et al., 2006). These approaches balance flexibility and performance, adapting to small, controlled datasets and large, noisy corpora.

However, while AA methods advance applications in politics, law enforcement, and cybersecurity, they also raise ethical and societal concerns. Risks include unauthorized surveillance, identity exposure, and bias against demographic groups, which may lead to systemic discrimination, misuse, or reputational harm (Juola, 2020; Lund et al., 2023; Alejo & Garrido-Merchan, 2023). Transparency and accountability are particularly challenging in high-stakes scenarios, such as legal or academic investigations (Boenninghoff et al., 2019), and AA systems can be weaponized to suppress dissent or manipulate public opinion (Shamsi et al., 2016).

Given these risks, the development and deployment of AA technologies must be grounded in strong ethical foundations, rigorous oversight, and a commitment to minimizing harm. In their research, Jobin et al. (2019),Benzie and Montasari (2023) highlight key ethical principles, including Transparency, Justice and Fairness, Non-Maleficence, Responsibility, and Privacy. Adapting directly from these principles, we establish guidelines specifically tailored for the stakeholders involved in the Software Development Life Cycle (SDLC) (Ruparelia, 2010) of AA tasks. These stakeholders range from system designers, project managers, developers, and testers to end-users, encompassing the three phases: Design and Planning, Development and Validation, and Deployment. Based on existing state-of-the-art (refer Sect. 2.1), our framework focuses on four key areas: privacy and data protection, fairness and non-discrimination, transparency and explainability, and consideration of broader societal impact. By addressing these aspects, we consolidate responsible AA practices in NLP to help guide all stakeholders in understanding and managing the inherent trade-offs in AA tasks. Figure 1 illustrates the overview of our framework (in Sect. 3) as a roadmap for implementing responsible AA

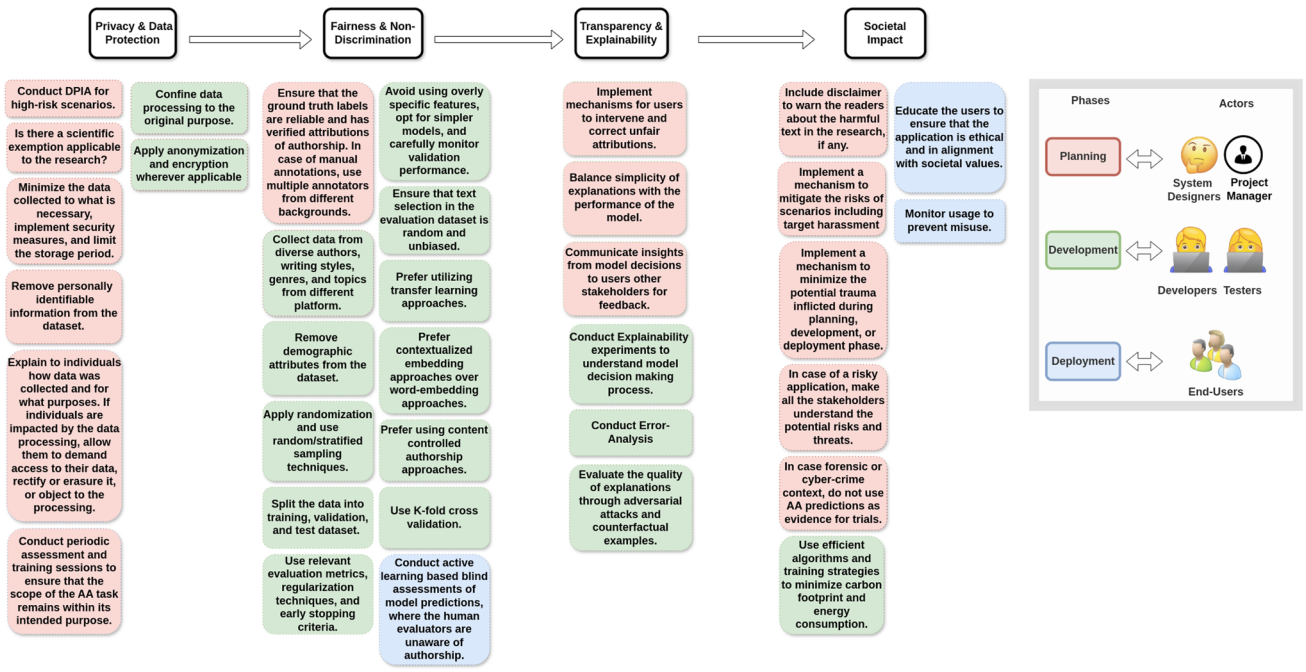

Fig. 1 The framework of responsible guidelines for the Authorship Attribution (AA) approaches in Natural Language Processing (NLP), encompassing aspects like Privacy & Data Protection, Fairness & Non-discrimination, Transparency & Explainability, and Social Impact. The guidelines are established throughout the Design and Planning, Development and Validation, and Deployment and Feedback phases of the Software Development Life Cycle (SDLC), involving System Designers, Project Managers, Developers, QA Testers, and End-Users as key stakeholders

practices. Finally, to demonstrate the practical application of our framework, we apply our guidelines to a recent AA study (Saxena et al., 2023a) that aims to identify potential human-trafficking vendors.[1]

Literature research

Foundation of the framework

Researchers have extensively explored the ethical debates surrounding AI development and deployment, identifying key challenges and principles. Notably, Jobin et al. (2019) mapped AI's most pressing ethical issues, emphasizing principles such as Transparency, Justice and Fairness, Non-Maleficence, Responsibility, and Privacy. In their research, Jobin et al. (2019),Benzie and Montasari (2023) focus on how these ethical principles apply to AI, ensuring its effective and responsible development. Transparency, for instance, requires that AA processes, algorithms, and data usage be clear to users and stakeholders. For example, when determining the author of a controversial document, the decision-making process must be explainable, with insights into how conclusions were drawn, which features were most influential, and what data was utilized. Justice and fairness are essential to prevent the perpetuation of biases or discrimination against certain groups. An AA model trained predominantly on texts from a particular demographic may unfairly attribute authorship to underrepresented groups. Non-maleficence underscores the need to avoid harm, such as the wrongful attribution of a document in forensic investigations, which could lead to false accusations or reputational damage. Responsibility requires all AA system stakeholders to be accountable for their outcomes. For instance, when a media organization uses AA to identify the author of an anonymous whistleblower article, it must ensure accuracy and handle findings with care to prevent harm. Finally, privacy is critical due to the risks of misusing personal data, particularly in cases where AA might expose an individual's identity against their will. Although several frameworks for ethical AI exist (Fjeld et al., 2020; Loi & Spielkamp, 2021; Loi et al., 2021; John Albert & Muller, 2022; Mollen, 2023; Charles Radclyffe, 2023), their direct application to AA remains limited, highlighting the need to address AA's unique challenges.

Our research adapts Jobin et al. (2019)'s principles for AA, emphasizing four pillars: Privacy and Data Protection, Fairness & Non-Discrimination, Transparency & Explainability, and Societal Impact. Our principle of Privacy and Data Protection, which aligns with the General Data Protection Regulation (GDPR) (Voigt & Bussche, 2017), directly addresses the privacy concerns highlighted by Jobin et al. (2019) by emphasizing data minimization, limited processing, purpose limitation, and the responsible handling of personal information (Klymenko et al., 2022; Habernal et al., 2023; Sousa & Kern, 2023). Fairness and Non-discrimination align with the principles of Justice and Fairness, ensuring that AA models do not perpetuate biases or discriminate against certain groups (Blodgett et al., 2020; Czarnowska et al., 2021; Halvani & Graner, 2021; Murauer & Specht, 2021a). The commitment to fairness also intersects with transparency and explainability, a cornerstone of our framework, ensuring that the processes and decisions in AA are clear, understandable, and accessible to stakeholders, thereby fostering trust and accountability (Escart'in et al., 2021). Furthermore, our principle of Societal Impact covers the ethical considerations of Non-Maleficence, focusing on the need to avoid risks, harm, and malicious use while including the environmental Impact of deploying large-scale AA systems and encouraging developers to evaluate the broader implications of AA technologies to strive for positive social outcomes (Hovy and Spruit, 2016; Bender et al., 2021). Finally, Responsibility is addressed by delegating clear guidelines to the various stakeholders involved in the AA development cycle, ensuring accountability and ethical practices. By adhering to these principles, we believe stakeholders can promote ethical practices that safeguard individuals' rights, enhance fairness and transparency, and ultimately contribute to a more just and equitable use of AA models.

Background on the ethical principles

The rapid advancement of ML has sparked debates about its benefits and risks, leading to the development of frameworks for responsible AI (Floridi & Cowls, 2019; Fjeld et al., 2020; Floridi et al., 2021). These frameworks promote fairness, accountability, and transparency (FAccT AI) (Simbeck, 2022; Laufer et al., 2022; Young et al., 2022), emphasizing socio-technical considerations (Dignum, 2019). Translating these principles into actionable guidelines for AA remains a critical gap. Our research builds on existing ethical AI frameworks (Mittelstadt, 2019; Felzmann et al., 2020; Prem, 2023) to address specific Ethical, Legal, and Social Issues (ELSI/ELSA) in AA, providing a questionnaire to guide stakeholders in navigating these challenges.[2]

[1] It should be noted that responsible data-sharing practices are beyond the scope of this research. However, we strongly encourage readers to follow the extensively detailed practices outlined by Gebru et al. (2021).

[2] This questionnaire is not a checklist but an awareness tool to guide discussions on ELSI/ELSA issues and offer ways to address them.

Privacy & data protection: In AA tasks within NLP, handling large volumes of textual data involves dealing with personal information, even if pseudonymized (Sjöberg, 2021). For instance, determining the authorship of emails or social media posts could inadvertently reveal sensitive personal information through writing styles or contextual clues (Sennewald et al., 2020). This underscores the need for strict adherence to data protection and privacy regulations, including the principles of lawfulness, purpose limitation, data minimization, and data security (De Terwangne, 2020a; OECD, 2013; Gellman, 2022). Privacy-by-design approaches must be embedded early in AA development to balance utility and individual rights (Tamo-Larrieux, 2018). Our research builds on these principles to propose guidelines that help assess AA applications, ensuring that they respect privacy and data protection norms while effectively serving their intended purposes.

Fairness & non-discrimination: The presence of biases in NLP models, including AA systems, is a critical issue that can lead to unfair or discriminatory outcomes (Shah et al., 2019; Chang et al., 2019; Blodgett et al., 2020; Hovy & Prabhumoye, 2021; Lalor et al., 2022). For example, suppose an AA model is trained predominantly on texts from a specific demographic. In these cases, it may unfairly attribute authorship when analyzing texts from underrepresented groups, thus perpetuating societal inequalities, reinforcing stereotypes, and compromising the fairness of the attribution process (Caliskan et al., 2017; Dev et al., 2022). Biases can enter the AA process at various stages, from data collection (Dixon et al., 2018) to feature selection (Bolukbasi et al., 2016; Ai et al., 2022), model training (Zafar et al., 2017), and evaluation (Delobelle et al., 2022). While existing research has explored biases in AA tasks (Bevendorff et al., 2019; Bischoff et al., 2020; Murauer & Specht, 2021a; Brad et al., 2021; Ai et al., 2022), our research draws from these insights to propose strategies for mitigating biases throughout the different phases of AA application development.

Transparency and explainability: Ensuring transparency in AA systems involves clarifying how decisions are made, which features are used, and how data is processed. This is crucial for building trust in AA systems, as it allows users to understand the rationale behind authorship conclusions. For instance, explaining how a specific writing style or linguistic pattern influenced an attribution decision can significantly enhance user trust and acceptance of the system's outcomes, especially within law enforcement applications. It ensures AA systems can be audited and justified, essential for maintaining integrity and fairness in algorithmic decision-making. The need for transparency and explainability has been widely discussed across various fields, including NLP (Chiticariu et al., 2015; Kim et al., 2020; Ma et al., 2020; Chen et al., 2020; Saxon et al., 2021; Ethayarajh & Jurafsky, 2021; Balkir et al., 2022b; Bhatt et al., 2022) and law and ethics (John-Mathews et al., 2022; Weinberg, 2022; Zhang et al., 2023). In the context of AA, transparency helps address two primary concerns: prospective transparency, which involves disclosing how data will be used and processed, and retrospective transparency, which focuses on explaining the outcomes of the algorithmic processes (Felzmann et al., 2019). Retrospective transparency is particularly relevant to the research on the explainability of algorithmic decision systems, ensuring that the steps leading to a decision can be understood and scrutinized (Ding et al., 2015; Boenninghoff et al., 2019; Bogdanova & Romanov, 2021; Theophilo et al., 2022; Kondyurin, 2022). This aspect is closely tied to accountability, as it verifies the system's compliance with ethical standards and responsiveness to user concerns (Procter et al., 2020; Qian et al., 2021; Rawal et al., 2021; Angelov et al., 2021; Balkir et al., 2022c). Although extensive research on transparency exists in the broader field of NLP (Shook et al., 2017; Tubella et al., 2019; Bogina et al., 2021; Angerschmid et al., 2022; Hacker et al., 2022), specific, actionable guidelines for implementing transparency in AA tasks are still lacking. Our research seeks to fill this gap by providing comprehensive guidelines that ensure transparent and fair data processing and decision-making, making AA systems accountable and trustworthy.

Societal impact: Despite its promises, the perceived accuracy and reliability of AA models make them susceptible to potential misuse and misrepresentation (Potthast et al., 2016; Suresh & Guttag, 2021; Zhai et al., 2022; Uchendu et al., 2023). These techniques can deceive or manipulate individuals by attributing text to different authors or falsely accusing someone of writing particular content, leading to significant ethical and legal ramifications. While AA holds potential for social welfare in fields such as forensic linguistics (Iqbal et al., 2008; Yang & Chow, 2014a; Wright & May, 2014; Fobbe, 2021), cybersecurity (Nirkhi & Dr.R.V.Dharaskar, 2013), and cybercrime detection (Zhang et al., 2019; Zheng et al., 2003; Kumar et al., 2020; Manolache et al., 2022; Saxena et al., 2023b), it shares similar risks and potential for misuse as other AI technologies (Juola, 2020). To mitigate these risks, it is crucial to understand the broader societal implications of AA, particularly the potential for harm and ethical breaches. The broader context of risks and harms associated with NLP technologies highlights the need for careful consideration of how AA models are designed, implemented, and deployed (Banko et al., 2020; Weidinger et al., 2021; Suresh & Guttag, 2021; Shmueli et al., 2021; Hovy and Spruit, 2016; Kirk et al., 2022; Haduong et al., 2023). Additionally, the environmental impact of AA models is an important societal aspect that must not be overlooked. The efficiency and robustness of trained AA models directly influence their carbon footprint and resource consumption (Saxena et al., 2023a). As the demand for immense and more complex models grows,

so does their environmental impact, raising concerns about sustainability and the responsible use of resources (Van Wynsberghe, 2021; Wu et al., 2022). Addressing these environmental considerations is crucial for ensuring that AA technologies are ethically sound and environmentally responsible. Drawing inspiration from these discussions, our research aims to provide comprehensive guidelines addressing AA applications' specific threats, environmental impacts, and broader society. By incorporating these considerations into each phase of the AA application's life cycle, we strive to promote responsible and ethical use of AA technologies, ensuring they contribute positively to society while minimizing risks and harms (Martin et al., 2020a; Madiega, 2021; Gerards et al., 2022; Tabassi, 2023). This integrated approach to understanding and mitigating the societal impacts of AA technologies, including environmental considerations, is essential for ensuring that these tools are used responsibly, safeguarding against misuse, and fostering trust and reliability in AA applications.

Stakeholders of the framework

To address the responsibility aspect highlighted by Jobin et al. (2019), our framework adopts the standard SDLC waterfall model, as outlined by Ruparelia (2010). This model's sequential and structured approach facilitates the clear identification of stakeholders and the assignment of responsibilities across each phase of software development. Key stakeholders include system designers, developers, project managers, quality assurance testers, and end-users. Each role ensures that AA technologies are developed and deployed ethically and responsibly. Our research categorizes these stakeholders into three distinct phases of the SDLC:

Design and planning phase: In this phase, we focus on data collection and processing, ensuring that designers and managers embed privacy considerations by incorporating ethical guidelines into the architecture, establishing a solid foundation for subsequent phases.

- Engaged in the early stages, system designers are responsible for embedding ethical principles into the architecture of the AA cycle. Their focus on ensuring privacy, fairness, and transparency can help lay a solid ethical foundation guiding the development process.
- In a collaborative effort with designers, managers oversee the entire development lifecycle. They ensure that ethical standards are upheld from the requirements gathering stage to deployment. Project managers are pivotal in coordinating between different teams, maintaining timelines, and, sometimes, holding all stakeholders accountable for ethical practices.

Development and validation phase: Developers translate ethical designs into functional models, addressing potential biases and ensuring compliance with privacy regulations. This phase involves not just training and evaluation, but also continuous monitoring to demonstrate the ongoing commitment to ethical development.

- Developers implement the algorithms defined by designers. They focus on creating systems that adhere to ethical guidelines, address potential biases, and protect user privacy.
- Testers evaluate the AA systems to ensure they function as intended and adhere to ethical guidelines. Their role is to identify and rectify any issues before deployment. Testers act as the final checkpoint, ensuring the system is ethically sound and ready for deployment.

Deployment and feedback phase: During the Deployment and Feedback phase, real-time applications are deployed. End-users play a critical role in this phase, providing feedback that is essential for refining the system and ensuring it adheres to ethical guidelines.

- End-users, such as forensic analysts, researchers, and media professionals, interact directly with the AA systems in their final form. Their feedback can be invaluable for assessing the system's performance and ethical impact. They help identify areas for improvement, ensuring the system remains aligned with ethical standards and continues to meet its intended purpose effectively.

By organizing stakeholders according to their roles in the SDLC with respect to their phases, we ensure a clear flow of responsibilities and accountability throughout the AA lifecycle. This structured approach is essential for integrating ethical principles at every stage, helping maintain AA technologies' integrity, trustworthiness, and societal value.

Proposed framework

Our research emphasizes integrating ethical principles into a practical framework, addressing the critical need for responsible AA system development. By employing the SDLC waterfall model, we systematically incorporate four key ethical principles-privacy and data protection, fairness and non-discrimination, transparency and explainability, and societal impact-throughout each phase of AA system development. This structured approach defines the roles and responsibilities of various stakeholders, including designers, project managers, developers, testers, and end-users, across the design and planning, development and validation,

and deployment phases. To operationalize these principles, we propose role-based guidelines and accompanying color-coded guiding questionnaire[3] designed to guide various SDLC stakeholders through the project phases: design and planning, development and validation, and deployment and feedback, with practical examples illustrated through a case study on AA applications in the context of cybercrime (Saxena et al., 2023a). The case study highlights how these guidelines address privacy risks, biases, transparency gaps, and societal harms, offering actionable recommendations. The appendix (Appendix A) includes details, explanations, and suggestions to balance space constraints and clarity. The questionnaire serves several purposes:

- **Operationalizing ethical principles:** The questionnaire transforms abstract concepts into practical, actionable steps by linking specific questions to the ethical guidelines discussed in each subsection.
- **Guiding stakeholders:** Each question is color-coded and tailored to the roles and responsibilities of stakeholders (e.g., system designers, project managers, developers, testers, and end-users), ensuring that ethical considerations are addressed at different AA development and deployment stages. The structured approach encourages stakeholders to reflect on their decisions and document their adherence to ethical principles, fostering transparency and accountability.[4]
- **Ensuring practical utility:** The detailed explanations and suggestions (in appendix Appendix A) highlight specific aspects of the framework, such as responsible and ethical practices, helping readers and practitioners quickly identify actionable insights and areas for improvement.

[3] The guidelines associated with Privacy and Data Protection are abbreviated as PD, Fairness and Non-Discrimination as FB (for Fairness and Biases), Transparency and Explainability as TE, and Societal Impact as SI in the questionnaire.

[4] Our role-specific guidelines represent a unique approach to integrating ethical principles into the AA development process. Although no existing literature specifically outlines these role-based guidelines, our assignments are carefully designed based on each stakeholder's established duties and responsibilities in the AA system lifecycle. This alignment ensures that each stakeholder's role adheres to the ethical principles of privacy, fairness, transparency, accountability, and societal impact, thereby supporting the responsible development of AA technologies.

Privacy & data protection

System designer and project managers

Privacy, a fundamental theme in ethical AI guidelines (Fjeld et al., 2020), remains a central issue, presenting challenges due to its elusive nature (Solove, 2005; Gasser, 2016; Nissim & Wood, 2018). In the evolving landscape of digital technologies and a data-driven society (Van Es & Schäfer, 2017), the focus of privacy research has shifted towards data privacy (Cukier & Mayer-Schoenberger, 2013), making data protection increasingly prominent (Lynskey, 2017). Regulations like the GDPR in the European Union (EU) (Regulation, 2018) are pivotal in addressing privacy concerns related to AA techniques (Saxena et al., 2023b).

To ensure compliance with privacy and data protection standards, system designers must conduct Data Protection Impact Assessments (DPIAs) to evaluate potential risks associated with data processing (Demetzou, 2019b; Martin et al., 2020b). Mandated by the GDPR, DPIAs help identify and mitigate privacy risks in high-risk scenarios by assessing how data is collected, stored, and used while implementing safeguards such as encryption and anonymization to protect against unauthorized access or data breaches. These assessments can also be considered exemptions for scientific research under GDPR Article 89 (Staunton et al., 2019). Designers can use various templates and tools to conduct DPIAs effectively (De Hert, 2012; Bieker et al., 2016; Mantelero, 2018; Voisin et al., 2020).

DPIAs further aid compliance with emerging AI regulations like the EU AI Act (Edwards, 2021), which categorizes AA systems based on risk levels (De Cooman, 2022; Hupont et al., 2023; Neuwirth, 2023). For instance, systems predicting behavior solely through profiling are prohibited (Article 5), but systems supporting human assessments may qualify for exemptions (Art. 5(d)). High-risk applications, such as law enforcement, require compliance documentation and registration with relevant authorities (Arts. 6(2-4), 49 AI Act).

Alongside conducting DPIAs, system designers are also responsible for implementing secure and compliant data storage practices within AA systems. Encryption and pseudonymization must be employed to safeguard data at rest and in transit, thereby reducing the risk of unauthorized access or breaches (Voigt & Von dem Bussche, 2017). To adhere to data minimization principles, designers should ensure data is retained only for the duration necessary to fulfill its intended purpose (Mondschein & Monda, 2019). Furthermore, robust authentication protocols and access controls must be integrated into the system architecture to restrict data access exclusively to authorized personnel (Kennedy & Millard, 2016). By embedding these measures into the system design,

compliance with privacy principles is reinforced, fostering trust and reducing risks throughout the AA system lifecycle.

Project managers are crucial in ensuring that AA systems align with privacy and data protection principles. Their responsibilities include facilitating stakeholder communication to ensure awareness and understanding of relevant privacy laws, such as the GDPR and the EU AI Act. They must coordinate with system designers and compliance officers to ensure that data handling practices are secure and compliant, protecting sensitive information from unauthorized access or breaches. Project managers are also responsible for documenting how data is collected, processed, and stored, ensuring that all practices are transparent and aligned with privacy principles (Demetzou, 2019a). They also uphold user rights, including access, rectification, objection, and erasure (Ausloos et al., 2019a). The GDPR permits automated decisions under certain conditions-such as explicit consent or legal requirements-provided safeguards are in place (Tamo-Larrieux, 2021). By overseeing these aspects, project managers ensure AA systems respect privacy, uphold high data protection standards, and foster trust, promoting responsible AI development.

PD1. Does the AA research/application under study involve a high level of risk, necessitating a Data Protection Impact Assessment (DPIA)? High-risk scenarios may include biometric identification, law enforcement, or justice system usage, etc.

PD2. Is there a scientific purpose or objective justifying exemptions from GDPR provisions?

PD3. What measures are in place to comply with privacy principles regarding data collection, processing, and storage in the AA system?

PD4. Are there specific protocols to protect sensitive/personal information from unauthorized access or breaches throughout the AA system?

PD5 . Is the information provided to individuals about data processing clear, complete, and correct?

PD6 . Is there periodic assessment and review to ensure ongoing relevance of data usage and AA application?

Developers and QA testers

Developers and testers play a critical role in ensuring that AA systems comply with GDPR data privacy principles, such as purpose limitation, data minimization, storage limitation, and data integrity and confidentiality (De Terwangne, 2020b; Bincoletto, 2020). They must design and evaluate AA systems with clear data processing policies that help users understand what data is being collected and how it will be used. They must also create user interfaces and consent mechanisms that provide straightforward controls over data usage.

to comply with the purpose limitation and data minimization, developers should process data only for clearly defined purposes, using anonymized or aggregated data where possible. They must implement storage limitation practices like automatic data deletion and robust security measures, including encryption, secure communication protocols, strict access controls, and threat detection systems. By embedding these privacy principles into the design and operation of AA systems, developers help protect user privacy, maintain trust, and ensure compliance with GDPR standards.

Throughout the AA development cycle, testers also play a vital role in adhering to GDPR data privacy principles (De Terwangne, 2020b; Bincoletto, 2020). Testers are responsible for verifying that the AA system complies with fairness, transparency, purpose limitation, data minimization, storage limitation, and data integrity and confidentiality. This includes validating that the system only processes data for purposes that have been clearly defined (purpose limitation), only relevant data is collected (data minimization), and storage limitation practices are effectively implemented by verifying automatic data removal protocols to ensure personal data is deleted when no longer necessary. Additionally, testers must assess the effectiveness of security measures like encryption and authorization controls to protect data integrity and confidentiality (Sweeney et al., 2015; Sion et al., 2021). By rigorously testing these aspects, testers help ensure AA systems comply with GDPR, protecting individual privacy rights throughout the system's lifecycle.

PD7. Is data processing confined to the original purpose for which it was collected? Is there periodic assessment and review to ensure ongoing relevance?

PD8. Have adequate safeguards like anonymization, encryption, data minimization, and security procedures been implemented to minimize risks and protect individual rights? Are these measures in line with guidelines from research and academic organizations, with ethical oversight?

Fairness and non-discrimination

Like other NLP applications, AA approaches are susceptible to unintended biases that can significantly impact their accuracy and fairness (Halvani & Graner, 2021; Murauer & Specht, 2021a). Numerous types of biases can impact the outcome of these tasks, which include label bias, selection bias, demographic and population bias, sampling bias, domain and genre bias, representation bias, model overfitting, evaluation bias, and user-interaction bias.

System designers

Label bias in NLP-based AA tasks arises when training data contains biased or incorrect authorship attributions stemming from systematic biases or human annotation errors (Shah et al., 2020). To mitigate this, system designers must ensure training data is reliable and its attributions

thoroughly validated and cross-checked (Shah et al., 2020). For human annotations, the designer should ensure that multiple domain experts independently assign authorship labels to the documents (Søgaard et al., 2014; Shah et al., 2020). Taking the consensus of these annotations can reduce the risks of individual biases or errors (Søgaard et al., 2014). Additionally, an active learning strategy (Zhang et al., 2022) can be employed to periodically review and update training data, incorporating corrections or new insights to maintain data quality and reliability.

FB1 . Is there a specific label or target for each data instance, and how were these labels obtained? For manually annotated data, please provide details about the number of annotators, their backgrounds, and any measures taken to mitigate label bias.

Developer

Selection bias arises when the dataset used for training systematically over-represents or under-represents certain groups or characteristics, resulting in a dataset that does not accurately reflect the target distribution (Cawley & Talbot, 2010). This can occur when non-random factors influence data selection, leading to skewed author distributions that affect the model's generalization. For example, if the data is collected from a specific platform or genre, it may not capture the diversity of authorship in the broader population. While semi-supervised approaches can help mitigate selection bias, they might not be effective if the gap between the source and target domains is too wide. In such cases, semi-supervised learning could increase bias rather than reduce it (Plank et al., 2014; Søgaard et al., 2014). To address this issue, developers should ensure that the dataset encompasses diverse authors, writing styles, genres, and topics that reflect the real-world distribution of texts. Combining data from multiple platforms helps reduce the risk of over-reliance on a single source (Shah et al., 2020).

Sampling bias, on the other hand, refers to any bias that occurs when the sample used for training does not adequately represent the entire population of interest (Hacker, 2018; Prabhu et al., 2019; Mehrabi et al., 2022). This can result from systematic selection issues, such as selection bias or random anomalies in the sampling process. While selection bias is a specific form of sampling bias, sampling bias more broadly encompasses any imbalance in the data distribution. This imbalance can prevent the model from generalizing well to new, unseen authors or writing styles. Developers should ensure that the training data reflects the population using random or stratified sampling (Prabhu et al., 2019). Although generalizing to every possible author is impossible, such methods can help achieve more balanced and representative model learning (Prabhu et al., 2019; Mehrabi et al., 2022).

Demographic & population bias stems from systematic patterns in the dataset that correlate with authors' demographic attributes or population characteristics (Mehrabi et al., 2022). These biases stem from variations in writing styles, vocabulary, or linguistic patterns associated with different demographic groups, such as age, gender, ethnicity, or geographical location. Consequently, the AA model gets swayed by these demographic cues, leading to unfair or inaccurate predictions that favor or penalize authors based on demographics rather than their writing style or content (Ai et al., 2022). While removing personally identifiable information (Dev et al., 2022) and demographic attributes (Bender & Friedman, 2018) helps, it does not necessarily correct the underlying demographic biases that influence model predictions. To mitigate these biases more effectively, developers must consider enriching the dataset with diverse writing samples across demographic groups or implementing algorithmic adjustments that compensate for potential biases in training data.

Domain & genre bias in AA relates to dataset patterns associated with text domains or genres, rather than writing style Julian et al. (2017). These biases result from language use and conventions specific to different domains or genres, leading to incorrect authorship attributions. To address this, the developer should ensure a balanced representation of authors from diverse domains and genres in the training dataset is crucial. Transfer learning, where the model is pre-trained on a broad, diverse dataset and then fine-tuned for the specific attribution task, can also help reduce domain and genre bias (Barlas & Stamatatos, 2021). Content-controlled authorship approaches can further assist in focusing on writing style by normalizing thematic content during training, disregarding domain-specific cues (Wegmann et al., 2022).

Overfitting refers to situations where the AA model performs exceptionally well on the training data but struggles to generalize to new texts (Schaffer, 1993; Lawrence & Giles, 2000). This leads to a high-variance problem, where the model becomes too tailored to the specific characteristics of the training data, making its predictions unreliable for unseen data. To mitigate overfitting, developers should split the dataset into a training set for model development, a validation set for tuning and selection, and a test set for evaluating model performance using k-fold cross-validation (Jabbar & Khan, 2015). Additionally, developers should avoid training on overly specific features, opt for simpler models when appropriate (Cawley & Talbot, 2010), use proper regularization and early stopping techniques (Cawley & Talbot, 2010), and carefully monitor validation performance to detect and address overfitting.

Underfitting occurs when a model is too simplistic to capture the complex patterns in authorship styles, leading to high bias in its predictions (Jabbar & Khan, 2015). In this case, the model fails to model the underlying relationships in the data, resulting in poor performance on both the training and test data. To avoid underfitting, developers should ensure not to use overly simplistic models (Wolfe & Caliskan, 2021), create ensemble models to leverage diverse perspectives (Jabbar & Khan, 2015), employ feature engineering techniques (Ali et al., 2022), fine-tune hyperparameters for optimal performance (Locatelli et al., 2022), apply regularization methods to avoid over-simplification (Wolfe & Caliskan, 2021; Locatelli et al., 2022), and employ cross-validation for robust generalization assessments (Jabbar & Khan, 2015).[5].

FB2. Does the training dataset sufficiently represent the entire authorship landscape, and what steps were taken to mitigate selection bias?

FB3. Are there correlations between authors in the dataset and specific demographic attributes or population characteristics?

FB4. Does the dataset cover multiple text genres or domains, and if not, what actions were taken to prevent biases related to domain and genre?

FB5. Is there a class imbalance in the dataset, and what measures are implemented to avoid over-representing certain authors? Describe any sampling strategies used to address sampling bias?

FB6. What feature extraction techniques were employed during training, and was fine-tuning performed on the target data?.

FB7. What precautions were taken to prevent overfitting and underfitting during model training?

QA Testers

Evaluation bias in NLP leads to incorrect conclusions about a model's capabilities (Suresh & Guttag, 2021). It can manifest when datasets lack diverse authorship, favoring frequent authors and affecting the performance of minority ones, which can compromise model reliability and performance. Therefore, the testers must ensure that the evaluation datasets mirror real-world scenarios with random and unbiased text selection (Cawley & Talbot, 2010). Choosing appropriate evaluation metrics aligned with the task's objectives is crucial for accurate assessments and informed deployment decisions (Powers, 2008; Hamalainen & Alnajjar, 2021; Mehrabi et al., 2022).

FB8. Do the chosen evaluation metrics align with the primary task objectives, and what insights can be provided about model generalization and robustness?

End users

User-interaction bias in AA arises from human involvement during model design, development, or evaluation, potentially leading to unintentional biases in training data and evaluation setups (Mehrabi et al., 2022). User interactions, like annotators' expectations or feedback (Hamalainen & Alnajjar, 2021), can influence the model's training and evaluation, skewing predictions. Therefore, deployers should conduct blind assessment, involving diverse users for broader perspectives, and using explainability tools for unbiased analysis (Srinivasan & Chander, 2021; Mehrabi et al., 2022; Hamalainen & Alnajjar, 2021).

FB9. Were independent blind assessments conducted by external evaluators, and can information about their backgrounds and diversity be provided?

Transparency and explainability

Transparency is crucial in AA tasks, particularly in compliance with GDPR (Ausloos et al., 2019b; Felzmann et al., 2019; Hacker & Passoth, 2020; Grünewald & Pallas, 2021) and the proposed AI Act (Hacker et al., 2022; Madiega, 2021). Clear disclosures aligned with regulations should use multi-channel communication to meet data subject expectations and ensure fairness, which intersects with transparency to prevent unjust and discriminatory practices (Bincoletto, 2020).

Explainable AI (XAI) techniques identify and mitigate unfairness issues in AA models (Theophilo et al., 2022; Ai et al., 2022), especially in sensitive applications like forensic text analysis (Solanke, 2022). Interactive XAI methods empower researchers to identify and correct unfair attributions while providing textual explanations for model decisions (Stevens et al., 2020; Alikhademi et al., 2021; Theophilo et al., 2022; Balkir et al., 2022c; Ai et al., 2022). Additionally, it can empower researchers and practitioners to promote transparency by intervening and correcting unfair

[5] We would like to clarify that in AA tasks, "bias" can refer to two distinct concepts: algorithmic bias, which pertains to the model's performance, and ethical bias, which relates to fairness and potential discrimination. Underfitting and overfitting are examples of algorithmic bias linked to the bias-variance trade-off (Belkin et al., 2019) This trade-off explains how models balance their ability to fit training data and generalize to new, unseen data. Underfitting leads to high bias, where a model is too simplistic, failing to capture patterns in the data. Overfitting results in high variance, where a model becomes overly complex and fits the training data too closely, reducing its ability to generalize. Both types of bias affect the model's accuracy but are distinct from ethical concerns about fairness or discrimination.

attributions on a case-by-case basis (Shrestha et al., 2017; Manolache et al., 2021; Huertas-Tato et al., 2022). Such approaches can effectively shed light on instances where potentially problematic correlations exert undue influence by facilitating the generation of textual explanations for predictions. Additionally, the AA tool must come with user instructions, including the provider's identity and contact details and information about the system's characteristics, capabilities, and limitations.

To implement transparency effectively, mechanisms should be in place that allow users to intervene and correct unfair attributions made by the AA system. Active learning strategies (Abbas et al., 2023), which involve engaging users or experts to provide feedback when incorrect attributions are detected, can provide an interface that allows users to flag or correct unfair attributions, enhancing the accuracy and fairness of the system over time. Another critical aspect of transparency involves communicating AA decisions to users and stakeholders. Establishing structured communication plans ensures that AA decisions are conveyed effectively to stakeholders, such as providing technical documentation and visualizations for system designers and developers and organizing training sessions for non-technical stakeholders. This approach helps ensure that all parties understand the decision-making processes and the implications of these decisions.

However, the pursuit of explainability is not devoid of challenges (Balkir et al., 2022a), such as striking a balance between simplicity and accuracy in explanations (Barredo Arrieta et al., 2019; Agarwal, 2020; Crook et al., 2023; Saeed & Omlin, 2023). Overly simplistic explanations may not fully capture the complexity of model logic, potentially leading to misunderstandings or even the introduction of new biases. Furthermore, evaluating the quality of explanations and their impact on the fairness of authorship analysis is an ongoing and evolving challenge (Ai et al., 2022; Saxena et al., 2023b). Continuous research and development are required to ensure that XAI techniques effectively aid the goal of fair and transparent AA tasks.

System designers and project managers

TE1. What steps are taken to balance simplicity and accuracy in the explanations provided by the AA system?

TE2. How are the AA decisions communicated to users and stakeholders?

Developer and QA testers

TE3. Are any explainability/interpretability experiments conducted to understand model decision-making process and predictions?

TE4. Are any error-analysis experiments conducted to understand false-positive and true-positive predictions?

TE5. Are there mechanisms for users to intervene and correct unfair attributions made by the AA system?

TE6. Is there a process for evaluating the quality of explanations provided by the AA system?

Societal impact

System designers

While AA approaches have valuable applications across various domains (Boukhaled & Ganascia, 2014; Enriquez et al., 2023; Barbon et al., 2017; Fobbe, 2021; Mateless et al., 2021; Keenan Jones, 2022; Saxena et al., 2023b), they also pose significant risks and harms that require careful consideration (Hovy and Spruit, 2016; Juola, 2020). In their research, Kirk et al. (2022) provide a comprehensive definition of harm in the context of AA, encompassing both content-related concerns and risks that can adversely affect the emotional, psychological, and physical well-being and safety of individuals, groups, or society. A prominent concern in AA is linked to privacy and confidentiality. When a trained AA model inadvertently reveals sensitive information about the author, leading to identity disclosure, reputational damage, or legal ramifications (Chaski, 2005b; Kirk et al., 2022; Banko et al., 2020). Another critical risk is the potential misuse and abuse of AA algorithms, which could enable malicious activities such as targeted harassment, social engineering, or the creation of deceptive content falsely attributed to others (Banko et al., 2020; Saxena et al., 2023b). Moreover, the utilization of AA approaches in forensics and cybercrime applications raises concerns about the potential trauma inflicted on individuals during the design, development, and deployment stages, as they may be exposed to the harmful nature of criminal content (Pyevich et al., 2003; Dubberley et al., 2015; Duran & Woodhams, 2022; Birze et al., 2023; Banko et al., 2020).

To minimize the risk of exposing individuals to harmful content, all stakeholders should clearly understand research objectives and potential threats (Renda et al., 2021; Kirk et al., 2022). Authors should prioritize subject protection

when publishing research and highlight potential risks associated with sensitive data, distancing research from harmful viewpoints, as suggested by Kirk et al. (2022). Collaborative teamwork can mitigate individual exposure to harmful content, fostering clear communication and providing essential support. Offering mental health and psychological support to team members dealing with harmful text is crucial to help them cope with potential challenges (Kirk et al., 2022).

To minimize potential misuse, ensuring the model's scope aligns with its intended purpose (Koops, 2021; Moraes et al., 2021) is crucial. Incorporating human oversight and intervention mechanisms is vital, allowing for thorough review and rejection of content with ethical concerns (Kirk et al., 2022). This human-in-the-loop approach safeguards against including harmful or ethically problematic content in the model's training data. Although humans can overlook things, introduce bias, or may lack the expertise to detect problematic content. Therefore, routine audits and updates are pivotal as proactive measures to anticipate and address potential ethical challenges. Regularly reviewing and refining the model ensures alignment with evolving ethical standards and societal expectations, underscoring the commitment to responsible AI development (Tabassi, 2023).

Using AA techniques as legal evidence in forensic and criminal cases faces challenges in meeting the admissibility standards set by the 1993 Daubert ruling (Gold et al., 1993). These standards require scientific testimony to include attributes like calculated error rates, empirical testing, standardized procedures, peer review, and acceptance within the scientific community. Traditional AA techniques often fall short of these criteria, limiting their use to investigations and preventing admission as legal evidence (Chaski, 1997; Yang & Chow, 2014b). To enhance AA's courtroom acceptance, Chaski (2005a),Frye and Wilson (2018) propose approaches that improve accuracy, establish error rates, evidence standards, and meet the stringent Daubert criteria (Gold et al., 1993). Additionally, all documentary evidence submitted to the court must undergo manual review to align with authentication and expert testimony standards (Howard, 2008).

SI1. Is there a disclaimer to alert readers to potentially harmful content in the research?

SI2. What measures are in place to minimize the potential trauma experienced by individuals during the design, development, and deployment stages? Are there regular check-ins among team members to ensure clear communication and support in maintaining a healthy and safe working environment? Is mental health and psychological support offered to team members dealing with harmful text?

SI3. Does the scope of the AA model align with its intended purpose to minimize potential misuse?

SI4. Does the AA processing encompass systematic and extensive automated processing that leads to decisions with legal or significant effects on individuals? Are measures in place to prevent identity disclosure, reputational damage, or legal ramifications?

SI5. Is there a mechanism to mitigate the risk of potential misuse and abuse, including scenarios involving targeted harassment, social engineering, or the creation of deceptive content falsely attributed to others?

Project manager

Human oversight and intervention mechanisms are crucial in sensitive AA applications to prevent ethical issues and potential harm. AA systems can inadvertently reveal sensitive information, misattribute content, or be misused for malicious purposes like targeted harassment or social engineering. Without oversight, these risks can lead to significant ethical breaches, privacy violations, or the misuse of data, which could harm individuals or groups, damage reputations, or lead to legal repercussions. Human oversight helps ensure that the AA model's outputs are aligned with ethical standards and societal norms, thus safeguarding against unintended negative consequences. The role of a project manager is to establish these oversight mechanisms and processes, ensuring that these are integrated into the project from the start. Additionally, they coordinate ethical review boards, schedule audits, and create feedback channels. Their primary responsibility is fostering a culture of ethical awareness and accountability, ensuring team members understand and uphold ethical guidelines. By embedding these practices, project managers safeguard against ethical risks and ensure the project achieves its technical and societal objectives.

SI6. Are there mechanisms for human oversight and intervention to review and reject content with ethical concerns?

Developer

Finally, minimizing the carbon footprint and energy consumption in training AA models is vital for aligning AI development with sustainability goals (Strubell et al., 2019; Bannour et al., 2021). Developers can take proactive steps by prioritizing efficient algorithms and training strategies that demand fewer computational resources, optimizing model architectures via knowledge distillation (Beyer et al., 2022; Panov et al., 2022), and exploring data-efficient techniques like transfer learning (Gupta et al., 2020; Barlas & Stamatatos, 2021; Hessenthaler et al., 2022; Silva et al., 2023; Saxena et al., 2023b). Monitoring and quantifying carbon emissions during AA model training with carbon tracking tools (Lacoste et al., 2019; Anthony et al., 2020) can provide insights for optimization and offsetting strategies.

SI7. Are efficient algorithms and training strategies given priority to minimize the carbon footprint and energy consumption? Is carbon tracking employed to monitor and quantify carbon emissions during Authorship Attribution model training, aiding in optimization and offsetting strategies?

End-users

End-users are responsible for ensuring AA tools are used ethically and in alignment with societal values, adhering to strict ethical codes for legitimate purposes like research, security, or intellectual property protection. Using AA applications for malicious purposes-such as targeted harassment, unauthorized surveillance, spreading misinformation, or falsely attributing content to harm others-can lead to significant ethical violations and legal consequences. Vigilance is essential to prevent abuse, uphold transparency, and respect privacy and individual rights. This includes being aware of potential abuse, actively preventing unethical behavior, and ensuring their use of the AA application respects individuals' rights and privacy. By following these ethical guidelines, end-users can help ensure AA applications contribute positively to society, upholding trust and integrity in their deployment.

SI8. How are end-users educated and held accountable to ensure that AA applications are used ethically and aligned with societal values? Are there mechanisms to monitor usage and prevent misuse, such as targeted harassment, unauthorized surveillance, or spreading misinformation?

Discussion & limitations

While comprehensive, our responsible AA guidelines have inherent limitations. Each AA task is a complex interplay of unique factors, such as the nature of the text data, the domain, and the intended use. Therefore, while our framework offers a structured approach, it may not encompass every specific nuance or consideration for every case. The framework's effectiveness in ensuring responsible AA needs rigorous testing for operationalizability. In other words, the real-world application of our framework to promote responsible practices is a crucial aspect currently missing. Variations in applications may necessitate prioritizing certain aspects of responsibility, while practical constraints-like data availability, resources, or context-may require trade-offs. Researchers must carefully weigh these trade-offs against the potential benefits of their models or approaches.

For instance, our case study (in appendix Appendix A) on using AA to identify potential human trafficking vendors revealed several practical takeaways and challenges. Implementing privacy measures, such as data masking and anonymization, was straightforward and effectively minimized the risk of individual identification, aligning with GDPR requirements. However, ensuring comprehensive data protection throughout the entire data lifecycle-including secure storage and handling-proved more challenging due to the complexity and variability of the data involved. Addressing fairness and non-discrimination presented another significant challenge. While mitigating algorithmic biases was manageable, ensuring a diverse dataset by collecting data from multiple platforms and demographics was resource-intensive and infeasible. Balancing transparency through explainability was also difficult. Although the authors generated some explanations for the model's decisions, these explanations were not always reliable. The complexity of the neural network made it difficult to debug and understand the underlying causes of certain outcomes. Finally, considering the social impact, such as providing mental health support to team members handling sensitive data, underscores the need for structured support systems and ethical oversight. Implementing these systems is not only important but essential. However, implementing such support can be challenging in a university-based research environment with a small team and limited resources.

These examples illustrate that the trade-offs in implementing responsible AA practices can vary significantly depending on the specific application. Given the diversity of AA applications and their unique challenges, providing

one-size-fits-all recommendations for every stakeholder and scenario is infeasible. Practitioners must recognize that while our guidelines offer a foundational approach, they may need to make context-specific adjustments and trade-offs to align with their specific application needs and ethical considerations. This emphasis on context-specificity is a key takeaway, as it provides flexibility while encouraging to apply the guidelines in a way that best suits the situation.

Ultimately, it is important to acknowledge that ensuring responsible AA is not a one-size-fits-all endeavor. While our framework provides valuable guidelines, practitioners should remain flexible and adapt it to their specific AA tasks' unique characteristics and requirements. Moreover, the continuous evaluation and refinement of the framework based on practical experiences and evolving ethical standards are not just important but crucial. This adaptability is necessary to address the ever-changing AA challenges and responsibilities landscape.

Conclusion

This research introduces a framework of a comprehensive set of responsible guidelines to address ELSI/ELSA considerations in AA within NLP. Central to our approach is a structured questionnaire that aids in navigating the complexities and trade-offs inherent in AA research, privacy and data protection, fairness and non-discrimination, transparency and explainability, and ultimately enabling the responsible development of AA tools for the greater societal good while upholding ethical standards. This framework aims to guide researchers and stakeholders at all stages of an AA tool's lifecycle, from initial design to deployment, helping identify and address potential ethical issues. Furthermore, we demonstrate the application of these guidelines through a case study (in appendix Appendix A) on recent sensitive AA research, providing a practical example of how they can inform the ethical assessment of AA projects. While our framework aims for comprehensiveness, it may not cover every unique AA application, necessitating rigorous real-world testing to ensure its effectiveness. Different contexts may require varying priorities and trade-offs, demanding flexibility and adaptation to specific AA task characteristics. Therefore, continuous evaluation and refinement of the framework, informed by practical experiences and evolving ethical standards, are essential.

Case study: practical utility of the proposed questionnaire

To illustrate the practical utility of our established guidelines, we provide a detailed questionnaire and case study inspired by a recent AA study (Saxena et al., 2023a). This study examines the application of AA in the high-stakes context of cybercrime, specifically targeting the identification of potential human trafficking vendors within the Backpage Escort Markets. By delving into this domain, we demonstrate how our guidelines can be effectively operationalized to address complex ethical considerations in sensitive scenarios.

To enhance the practical application of our framework, we have designed a comprehensive, color-coded questionnaire. This tool supports stakeholders throughout the various phases of the Software Development Life Cycle (SDLC), including design and planning, development and validation, and deployment and feedback and associated stakeholders- system designers and project managers, developers and testers, and end-users -ensuring ethical considerations are systematically addressed at each stage of AA development and deployment.

Each question is color-coded and aligned with one of the core ethical principles: Privacy and Data Protection (PD), Fairness and Non-Discrimination (FB, abbreviated for Fairness and Biases), Transparency and Explainability (TE), and Societal Impact (SI). This structured format encourages stakeholders to reflect on their decisions, document their adherence to ethical standards, and maintain transparency and accountability throughout the process.

The questionnaire's structure directly corresponds to the framework outlined in the main manuscript, ensuring a direct integration between theoretical guidelines and their real-world application. By offering role-specific and phase-specific prompts, the questionnaire facilitates ethical reflection and supports responsible decision-making. This approach is critical for ensuring AA systems are responsibly designed, implemented, and deployed, especially in domains with profound societal implications.

Privacy & data protection

System designer and project managers

PD1. Does the AA research/application under study involve a high level of risk, necessitating a Data Protection Impact Assessment (DPIA)? High-risk scenarios may include biometric identification, law enforcement, or justice system usage, etc. The authors do not explicitly state whether a DPIA was conducted.
Suggestion: Given the application's high-risk nature, we suggest conducting a DPIA to identify and mitigate risks associated with data processing proactively. This would align with GDPR requirements and provide a structured approach to handling data responsibly. DPIAs are essential for evaluating the potential impact of data processing activities, particularly when dealing with high-risk scenarios like profiling individuals based on writing styles or patterns. Implementing a DPIA can help anticipate privacy issues, establish controls to mitigate risks, and ensure ongoing compliance with regulatory standards.

PD2. Is there a scientific purpose or objective justifying exemptions from GDPR provisions? The research aims to combat human trafficking, a significant societal issue, which could be considered a legitimate scientific purpose. The research provides these details under Section 8, Privacy Considerations and Potential Risk paragraph.

PD3. What measures are in place to comply with privacy principles regarding data collection, processing, and storage in the AA system? While Appendix sections A.2.3 and A.2.4 provide detailed information about data collection and processing, there is a lack of information regarding data storage practices.
Suggestion: To fully comply with privacy principles, it is important to include comprehensive details on data storage practices in addition to data collection and processing. The documentation should also address how data is stored, including security measures, retention policies, and access controls.

PD4. Are there specific protocols to protect sensitive/personal information from unauthorized access or breaches throughout the AA system? The personal data within the dataset, including phone numbers and email addresses, is masked, minimizing the risk of individual identification. The research provides these details under Section 8, Privacy Considerations and Potential Risk paragraph. However, while personal identifiers are masked, the inherent nature of the AA task in NLP involves linking writing styles (as a unique signature in the raw data) to individuals. The case study does not explicitly address how to mitigate this potential privacy concern.
Suggestion: To mitigate these risks, federated learning can be used to train models locally, reducing centralized data exposure. Incorporating consent mechanisms and allowing individuals to opt out of data use would further enhance privacy protection.

PD5 . Is the information provided to individuals about data processing clear, complete, and correct? There is no direct mention of how or if individuals are informed about data processing. For GDPR compliance and ethical considerations, ensuring transparency about data use is critical, allowing individuals to understand and, where applicable, contest the use of their data.
Suggestion: To comply with GDPR and uphold ethical standards, providing individuals with clear, complete, and accurate information about how their data is being processed is essential. This includes detailing the purpose of data collection, the specific data being used, how long it will be retained, and individuals' rights over their data. Implementing transparent communication practices, such as privacy notices or consent forms, ensures that individuals are informed and can contest or opt out of data processing if necessary. Regularly updating these communications to reflect any changes in data processing practices is also recommended.

PD6 . Is there periodic assessment and review to ensure ongoing relevance of data usage and AA application? No such details are provided in the chosen case study.
Suggestion: The project manager must conduct regular assessments and training sessions to ensure data usage remains relevant and compliant throughout the AA application cycle. These periodic reviews will help maintain the integrity of the data, adapt to any changes in the context or requirements, and ensure ongoing alignment with privacy and data protection standards.

Developers and QA testers

PD7. Is data processing confined to the original purpose for which it was collected? Is there periodic assessment and review to ensure ongoing relevance? Yes, the dataset is specifically compiled for AA to identify human trafficking operations, aligning with the original data collection purpose. However, periodic assessments and reviews are not explicitly mentioned yet could be essential for maintaining data relevance and compliance.
Suggestion: It is essential to ensure that data processing aligns with the original purpose of data collection. This can help design the system to restrict data use to its intended purpose and set up mechanisms for periodic assessments and reviews to maintain relevance and compliance. Furthermore, the project manager should conduct periodic assessments and training to maintain data relevance and compliance throughout the AA application cycle.

PD8. Have adequate safeguards like anonymization, encryption, data minimization, and security procedures been implemented to minimize risks and protect individual rights? Are these measures in line with guidelines from research and academic organizations, with ethical oversight? The research outlines anonymization and data minimization measures, such as masking personal details. However, the detailed security procedures are not specified, the described practices indicate an effort to align with ethical research guidelines. Implementing comprehensive security procedures and seeking ethical oversight would further enhance data protection. The research provides these details under Section 8, Privacy Considerations and Potential Risk paragraph.
Suggestion: To enhance data protection, comprehensive security procedures with strict access controls must be implemented alongside the existing anonymization and data minimization practices. Additionally, obtaining ethical oversight and regular audits is crucial. These audits play a significant role in ensuring that all privacy considerations are adequately addressed, instilling a sense of security and confidence in the data protection measures.

Fairness and non-discrimination

System Designers

FB1 . Is there a specific label or target for each data instance, and how were these labels obtained? For manually annotated data, please provide details about the number of annotators, their backgrounds, and any measures taken to mitigate label bias. Vendor labels are generated based on phone number connections among ads, not manual annotation. This approach aims to reduce bias associated with manual labeling, but the authors do not discuss measures to mitigate bias inherent in the data collection or algorithmic processing. The research provides these details under Section 8, Privacy Considerations and Potential Risk paragraph.

Developer

FB2. Does the training dataset sufficiently represent the entire authorship landscape, and what steps were taken to mitigate selection bias? The dataset comprises a significant number of ads from the Backpage escort market from 14 states and 41 cities of the U.S., aiming to capture a broad spectrum of authorship styles linked to potential human trafficking. The research provides these details under Section 3, Dataset and Appendix A.4, Datasheet.

FB3. Are there correlations between authors in the dataset and specific demographic attributes or population characteristics? The research does not provide an analysis of the authors' demographic attributes or population characteristics, potentially missing insights into bias related to these factors.
Suggestion: To gain a deeper understanding of potential biases in the dataset, it would be beneficial to investigate the correlations between the authors and specific demographic attributes or population characteristics. By identifying these correlations, the research can assess whether the model performs equitably across different demographic groups, thereby ensuring fairness and reducing the risk of bias. This analysis can also provide valuable insights into how representative the dataset is of the broader population and inform adjustments to improve the model's overall effectiveness.

FB4. Does the dataset cover multiple text genres or domains, and if not, what actions were taken to prevent biases related to domain and genre? The dataset focuses on text escort advertisements from a single source (Backpage), indicating a specific domain focus. The document does not describe actions taken to address potential biases arising from this focus, such as incorporating or analyzing ads from varied platforms or text genres to enhance generalizability.
Suggestion: To address potential domain and genre biases, it would be useful to expand the dataset to include text from multiple sources and platforms, covering various genres beyond the Backpage escort advertisements. This could incorporate data from online escort websites or posts.

FB5. Is there a class imbalance in the dataset, and what measures were implemented to avoid over-representing certain authors? Describe any sampling strategies used to address potential sampling bias? The research mentions the distribution of ads per vendor but does not detail specific measures to address potential class imbalance or sampling bias. To consider class imbalance, the authors emphasize using Macro-F1 and R-Precision metrics for evaluation.
Suggestion: Strategies like stratified sampling or synthetic data generation could help mitigate these issues, ensuring a balanced representation of different authors. The research provides these details under Section 3, Dataset.

FB6. What feature extraction techniques were employed during training, and was fine-tuning performed on the target data? The research employs various contextualized (transformers-based) models for feature extraction, focusing on style representations extracted from classified ads. Fine-tuning on the target dataset (IDTraffickers) is conducted to adapt the model to the specific domain, indicating an effort to capture domain-specific authorship styles effectively. The research provides these details under Section 4, Experimental Setup, and Section 5, Results.

FB7. What precautions were taken to prevent overfitting and underfitting during model training? The research outlines the model training process, indicating the data-split of 0.75:0.05:0.20 for training, validation, and test datasets. No explicit precautions against overfitting and underfitting are discussed in the research. Measures like cross-validation, regularization, and early stopping are typically employed to address these issues, ensuring model robustness. The research provides these details under Appendix A.4, Datasheet.

QA testers

FB8. Do the chosen evaluation metrics align with the primary task objectives, and what insights can be provided about model generalization and robustness? The research highlights the use of macro-F1 and mean r-precision scores for evaluating model performance, focusing on identifying authorship patterns across different vendors. These metrics are relevant to the task objectives, but additional analysis on model generalization and robustness, possibly through external validation or cross-domain testing, would provide deeper insights. The research provides these details under Section 4, Experimental Setup and Section 5, Results.

End users

FB9. Were independent blind assessments conducted by external evaluators, and can information about their backgrounds and diversity be provided? Independent blind assessments by external evaluators are not mentioned in the research.
Suggestions: Such assessments could enhance the credibility of the research findings. Providing information about the backgrounds and diversity of these evaluators would further strengthen the research, ensuring that a wide range of perspectives is considered. This approach can help identify potential biases in the model's evaluation and promote a more thorough and unbiased assessment of its performance.

Transparency and explainability

System designers and project managers

TE1. What steps are taken to balance simplicity and accuracy in the explanations provided by the AA system? The research utilizes local and global feature attribution techniques for qualitative analysis. These techniques allow the AA system to highlight the most relevant features contributing to a decision, simplifying complex model logic without compromising accuracy.

TE2. How are the AA decisions communicated to users and stakeholders? The chosen case study provides no specific details on communication methods for AA decisions. This is expected in a research environment where all participants often share stakeholder responsibilities.
Suggestion: In a typical SDLC, it is crucial to establish clear roles and responsibilities for communication. This ensures a secure and organized communication process and helps avoid any potential misunderstandings. To address this, project managers can implement a structured communication plan that defines how AA decisions are conveyed to different stakeholders at each phase of the SDLC. For example, system designers and developers can provide technical documentation and visualizations to explain decision-making processes. At the same time, end-user training sessions can help non-technical stakeholders understand the implications of these decisions. Regular updates and feedback loops can ensure that stakeholders remain informed and engaged throughout the development and deployment of the AA system.

Developer and QA testers

TE3. Are any explainability/interpretability experiments conducted to understand model decision-making process and predictions? The research utilizes local and global feature attribution techniques for qualitative analysis, indicating an effort toward explainability. The research provides these details under Section 5.3, Qualitative Analysis.

TE4. Are any error-analysis experiments conducted to understand false-positive and true-positive predictions? The research provides a systematic error analysis of true and false-positive predictions to understand the underlying causes of these outcomes. The research provides these details under Section 5.3, Qualitative Analysis.

TE5. Are there mechanisms for users to intervene and correct unfair attributions made by the AA system? No such details are provided in the chosen case study.
Suggestion: To address this gap, the AA system could use active learning strategies (Abbas et al, 2023) to engage users or experts when incorrect attributions arise, allowing user input to improve accuracy. An interface for users to flag or correct unfair attributions can help ensure ongoing fairness and reliability, creating a feedback loop that enhances the system over time.

TE6. Is there a process for evaluating the quality of explanations provided by the AA system? No such details are provided in the chosen case study.
Suggestion: To assess the quality of explanations in the AA system, robustness checks can be conducted using techniques such as adversarial attacks (Zhang et al, 2020) or counterfactual examples (Stepin et al, 2021). These methods evaluate how well the explanations withstand intentional disturbances or hypothetical changes, ensuring the system's explanations are dependable and resistant to manipulation. Implementing these checks helps maintain the integrity and trustworthiness of the AA system's outputs.

Societal impact

System designers

SI1. Is there a disclaimer to alert readers to potentially harmful content in the research? The research does not mention a specific disclaimer regarding potentially harmful content. Given the topic's sensitive nature, incorporating a disclaimer would be prudent to inform readers and mitigate any distress caused by the content.
Suggestion: Given the sensitive nature of AA research, we advise including a clear disclaimer alerting readers to potentially harmful content. This disclaimer would help prepare readers and mitigate any distress caused by exposure to sensitive information or topics discussed in the research.

SI2. What measures are in place to minimize the potential trauma experienced by individuals during the design, development, and deployment stages? Are there regular check-ins amongst team members to ensure clear communication and support maintaining a healthy and safe working environment? Is mental health and psychological support offered to team members dealing with harmful text? Measures to minimize potential trauma or support team members are not mentioned. Given the sensitive nature of the data, establishing mental health support and regular check-ins can help maintain a healthy working environment.
Suggestion: To address the potential trauma associated with handling sensitive content, it is crucial to implement measures such as providing access to mental health and psychological support services for all team members. Regular check-ins and open communication channels should be established to monitor individuals' well-being and promptly address any concerns. Additionally, incorporating training on resilience and coping strategies, as well as rotating tasks to limit prolonged exposure to harmful content, can further support maintaining a healthy and safe working environment.

SI3. Does the scope of the AA model align with its intended purpose to minimize potential misuse? The model's development and application specifically aim to identify potential human trafficking operations, suggesting alignment with the intended purpose. However, outlining explicit use cases and restrictions could further minimize potential misuse.
Suggestion: To minimize potential misuse of the AA model, we recommend clearly defining its scope and intended purpose by specifying explicit use cases, limitations, and restrictions. This can help ensure the model is used ethically and aligns with its intended goals, reducing the risk of misuse.

SI4. Does the AA processing encompass systematic and extensive automated processing that leads to decisions with legal or significant effects on individuals? Are measures in place to prevent identity disclosure, reputational damage, or legal ramifications? While the research involves automated processing to identify potential human trafficking operations, specific measures to prevent identity disclosure and reputational damage are mentioned, such as data anonymization and personal information masking. Furthermore, the authors emphasize that law enforcement agencies and researchers should not solely depend on our analysis as evidence for criminal prosecution. However, the research does not detail safeguards against potential legal consequences for individuals incorrectly identified. The research provides these details under Section 8, Privacy Considerations and Potential Risk, and Legal Impact paragraphs.
Suggestion: To enhance safeguards, we recommend establishing protocols to verify AA results before any legal action is taken and issuing disclaimers on their use as evidence. If AA results are to be used for prosecution, the methods and findings must be scientifically validated, empirically tested, and presented by qualified experts who can clearly explain them to jurors and withstand cross-examination (Howard, 2008).

SI5. Is there a mechanism to mitigate the risk of potential misuse and abuse, including scenarios involving targeted harassment, social engineering, or the creation of deceptive content falsely attributed to others? While the research outlines privacy-preserving measures under Section 8, Privacy Considerations and Potential Risk paragraph, it does not specifically address mechanisms to mitigate misuse and abuse, such as targeted harassment. Implementing comprehensive ethical guidelines and employing restrictions is essential for minimizing such risks.
Suggestion: Implementing comprehensive restrictions such as user access controls, monitoring systems, and reporting protocols can help prevent misuse and abuse, including targeted harassment, social engineering, and the creation of deceptive content. Incorporating a robust consent framework, where stakeholders are crucial in being informed of potential risks and agreeing to ethical usage, can also be valuable. Regular training sessions for users on ethical considerations and potential risks, combined with periodic audits and updates to the system, would ensure ongoing alignment with ethical standards and reduce the likelihood of misuse. Finally, establishing a clear process for detecting and responding to misuse can provide a practical response to ethical breaches.

Project manager

SI6. Are there mechanisms for human oversight and intervention to review and reject content with ethical concerns? The research does not detail mechanisms for human oversight and intervention.
Suggestion: To ensure ethical integrity, it is recommended to implement mechanisms for human oversight and intervention. This could include establishing an ethical review board, regular audits, and a pre-release content review process, allowing for identifying and rejecting content with ethical concerns. These measures will help prevent misuse and uphold ethical standards.

Developer

SI7. Are efficient algorithms and training strategies given priority to minimize the carbon footprint and energy consumption? Is carbon tracking employed to monitor and quantify carbon emissions during Authorship Attribution model training, aiding in optimization and offsetting strategies? The environmental impact of the research, including carbon footprint and energy consumption, is not discussed. Prioritizing efficient algorithms and employing carbon tracking would contribute to sustainable research practices, aligning with broader environmental responsibility goals.
Suggestion: Given the potential impact of AI development on the environment, it is essential to prioritize efficient algorithms and training strategies that reduce the carbon footprint and energy consumption. Incorporating carbon tracking tools during AA model training can provide valuable data for monitoring and quantifying emissions, enabling targeted optimization and offsetting strategies. These steps can help minimize the environmental impact of AI development, which is of urgent importance.

End-users

SI8. How are end-users educated and held accountable to ensure that AA applications are used ethically and aligned with societal values? Are there mechanisms to monitor usage and prevent misuse, such as targeted harassment, unauthorized surveillance, or spreading misinformation? There are no such details provided in the chosen case study.
Suggestion: To ensure the ethical use of AA applications, end-users should receive mandatory training on ethical guidelines and societal values. Additionally, implementing monitoring systems to track usage and clear policies and consequences for misuse can help prevent unethical behavior such as targeted harassment, unauthorized surveillance, or spreading misinformation. Regular audits and user compliance checks can further enhance accountability.

In summary, this section shows how the proposed questionnaire can be used as a practical tool for translating ethical guidelines into actionable steps throughout the AA lifecycle. The case study illustrates the complexities of applying these principles and emphasizes the need for flexibility in addressing domain-specific challenges, such as balancing data privacy with system utility and ensuring fairness despite limited resources. Importantly, the structured questionnaire approach encourages stakeholder reflection, accountability, and continuous improvement. It reinforces the alignment of technical decisions with ethical priorities. This exercise highlights the significance of integrating ethical considerations early and iteratively, acknowledging that responsible AA development is an evolving process influenced by context-specific trade-offs and emerging challenges.

Author contributions All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by *Vageesh Saxena* and *Aurelia Tamò-Larrieux*. The first draft of the manuscript was written by *Vageesh Saxena*, and all authors commented on previous versions. All authors read and approved the final manuscript.

Funding This research is supported by the Sector Plan Digital Legal Studies of the Dutch Ministry of Education, Culture, and Science. The Open-access funding will be provided by Maastricht University under the 100% APC discount.

Declarations

Conflict of interest The authors have no Conflict of interest to declare relevant to this article's content.

References

Abbas, S., Alsubai, S., Sampedro, G. A., et al. (2023). Active learning for news article's authorship identification. *IEEE Access, 11*, 98415–98426. https://doi.org/10.1109/ACCESS.2023.3310813

Agarwal, S. (2020). Trade-offs between fairness, interpretability, and privacy in machine learning. https://api.semanticscholar.org/CorpusID:229087464

Ai, B., Wang, Y., & Tan, Y., et al. (2022). Whodunit? learning to contrast for authorship attribution. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online only, pp 1142–1157, https://aclanthology.org/2022.aacl-main.84

Alejo, José G., & Sison RGBMarco Tulio Daza, Garrido-Merchán EC,. (2023). Chatgpt: More than a "weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (hcai) perspective. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2023.2225931

Ali, I., Mughal, N., & Khand, Z.H., et al. (2022). Mehran University Research Journal Of Engineering & Technology 41(1):65–79. https://search.informit.org/doi/10.3316/informit.263278216314684

Alikhademi, K., Richardson, B., & Drobina, E., et al. (2021). Can explainable ai explain unfairness? a framework for evaluating explainable ai. arXiv:2106.07483

Altheneyan, A. S., & Menai, M. E. B. (2014). Naïve bayes classifiers for authorship attribution of arabic texts. *Journal of King Saud University-Computer and Information Sciences, 26*(4), 473–484.

Angelov, P., Soares, E. V., Jiang, et al. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining Knowl Discov, 11*,. https://doi.org/10.1002/widm.1424

Angerschmid, A., Zhou, J., Theuermann, K., et al. (2022). Fairness and explanation in ai-informed decision making. *Machine Learning and Knowledge Extraction, 4*(2), 556–579.

Anthony, L.F.W., Kanding, B., & Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv:2007.03051

Ausloos, J., Mahieu, R., & Veale, M. (2019). Getting data subject rights right: A submission to the european data protection board from international data rights academics, to inform regulatory guidance. *JIPITEC-Journal of Intellectual Property, Information Technology and E-Commerce Law, 10*(3), 283–309.

Ausloos, J., et al. (2019b). Gdpr transparency as a research method. SSRN Electronic Journal, May pp 1–23

Balkir, E., Kiritchenko, S., & Nejadgholi, I., et al. (2022a). Challenges in applying explainability methods to improve the fairness of NLP models. In Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022). Association for Computational Linguistics, Seattle, U.S.A., pp 80–92, https://doi.org/10.18653/v1/2022.trustnlp-1.8, https://aclanthology.org/2022.trustnlp-1.8

Balkir, E., Kiritchenko, S., & Nejadgholi, I., et al. (2022b). Challenges in applying explainability methods to improve the fairness of nlp models. arXiv preprint arXiv:2206.03945

Balkir, E., Kiritchenko, S., & Nejadgholi, I., et al. (2022c). Challenges in applying explainability methods to improve the fairness of nlp models. ArXiv abs/2206.03945

Banko, M., MacKeen, B., & Ray, L. (2020). A unified taxonomy of harmful content. In: Proceedings of the Fourth Workshop on Online Abuse and Harms. Association for Computational Linguistics, Online, pp 125–137, https://doi.org/10.18653/v1/2020.alw-1.16, https://aclanthology.org/2020.alw-1.16

Bannour, N., Ghannay, S., & Névéol, A., et al. (2021). Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing. Association for Computational Linguistics, Virtual, pp 11–21, https://doi.org/10.18653/v1/2021.sustainlp-1.2, https://aclanthology.org/2021.sustainlp-1.2

Barbon, S., Igawa, R. A., & Bogaz Zarpelão, B. (2017). Authorship verification applied to detection of compromised accounts on online social networks. *Multimedia Tools and Applications, 76*(3), 3213–3233. https://doi.org/10.1007/s11042-016-3899-8

Barlas, G., & Stamatatos, E. (2021). A transfer learning approach to cross-domain authorship attribution. *Evolving Systems, 12*(3), 625–643.

Barredo Arrieta, A. (2019). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*. https://doi.org/10.1016/j.inffus.2019.12.012

Belkin, M., Hsu, D., Ma, S., et al. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849–15854.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604.

Bender, E.M., Gebru, T., & McMillan-Major, A., et al. (2021). On the dangers of stochastic parrots: Can language models be too big??? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. association for computing machinery, New York, NY, USA, FAccT '21, p 610–623, https://doi.org/10.1145/3442188.3445922,

Benzie, A., & Montasari, R. (2023). *Bias, privacy and mistrust: Considering the ethical challenges of artificial intelligence* (pp. 1–14). Springer Nature Switzerland.

Bevendorff, J., Hagen, M., & Stein, B., et al. (2019). Bias analysis and mitigation in the evaluation of authorship verification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 6301–6306, https://doi.org/10.18653/v1/P19-1634, https://aclanthology.org/P19-1634

Bevendorff, J., Chulvi, B., & Fersini, E., et al. (2022). Overview of pan 2022: Authorship verification, profiling irony, stereotype spreaders, style change detection. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings. Springer-Verlag, Berlin, Heidelberg, p 382–394, https://doi.org/10.1007/978-3-031-13643-6_24,

Beyer, L., Zhai, X., & Royer, A., et al. (2022). Knowledge distillation: A good teacher is patient and consistent. arXiv:2106.05237

Bhatt, S., Dev, S., & Talukdar, P.P., et al. (2022). Re-contextualizing fairness in nlp: The case of india. In: AACL

Bieker, F., Friedewald, M., & Hansen, M., et al. (2016). A process for data protection impact assessment under the european general data protection regulation. In: Privacy Technologies and Policy: 4th Annual Privacy Forum, APF 2016, Frankfurt/Main, Germany, September 7-8, 2016, Proceedings 4, Springer, pp 21–37

Bincoletto, G. (2020). Edpb guidelines 4/2019 on data protection by design and by default. *Eur Data Prot L Rev, 6*, 574.

Birze, A., Regehr, K., & Regehr, C. (2023). Workplace trauma in a digital age: The impact of video evidence of violent crime on criminal justice professionals. *Journal of interpersonal violence, 38*(1–2), 1654–1689.

Bischoff, S., Deckers, N., & Schliebs, M., et al. (2020). The importance of suppressing domain style in authorship analysis. ArXiv abs/2005.14714

Blodgett, S.L., Barocas, S., & au2, H.D.I., et al. (2020). Language (technology) is power: A critical survey of "bias" in nlp. arXiv:2005.14050

Boenninghoff, B., Hessler, S., & Kolossa, D., et al. (2019). Explainable authorship verification in social media via attention-based similarity learning. In: 2019 IEEE International Conference on Big Data (Big Data), IEEE, pp 36–45

Bogdanova, A., & Romanov, V. (2021). Explainable source code authorship attribution algorithm. *Journal of Physics: Conference Series, 2134*(1), 012011. https://doi.org/10.1088/1742-6596/2134/1/012011

Bogina, V., Hartman, A., Kuflik, T., et al. (2021). Educating software and ai stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education., 1*, 1–26.

Bolukbasi, T., Chang, K.W., & Zou, J., et al. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv:1607.06520

Boukhaled, M. A., & Ganascia, J. G. (2014). Probabilistic anomaly detection method for authorship verification. In L. Besacier, A. H. Dediu, & C. Martín-Vide (Eds.), *Statistical Language and Speech Processing* (pp. 211–219). Springer International Publishing.

Brad, F., Manolache, A., & Burceanu, E., et al. (2021). Rethinking the authorship verification experimental setups. In: Conference on Empirical Methods in Natural Language Processing

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230, https://www.science.org/doi/abs/10.1126/science.aal4230

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research, 11*, 2079–2107.

Chang, K.W., Prabhakaran, V., & Ordonez, V. (2019). Bias and fairness in natural language processing. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts. Association for Computational Linguistics, Hong Kong, China, https://aclanthology.org/D19-2004

Charles Radclyffe RHWMafalda Ribeiro (2023) The assessment list for trustworthy artificial intelligence: A review and recommendations — frontiersin.org. https://www.frontiersin.org/articles/10.3389/frai.2023.1020592/full, [Accessed 19-Jul-2023]

Chaski, C. (1997). Who wrote it? steps toward a science of authorship identification. *National Institute of Justice Journal, 233*(233), 15–22.

Chaski, C.E. (2005a). Who's at the keyboard? authorship attribution in digital evidence investigations. Int J Digit EVid 4. https://api.semanticscholar.org/CorpusID:12767441

Chaski, C. E. (2005). Who's at the keyboard? authorship attribution in digital evidence investigations. *International Journal of Digital Evidence, 4*(1), 1–13.

Chen, J., Berlot-Attwell, I., & Hossain, S., et al. (2020). Exploring text specific and blackbox fairness algorithms in multimodal clinical nlp. ArXiv abs/2011.09625

Chiticariu, L., Li, Y., & Reiss, F. (2015). Transparent machine learning for information extraction: state-of-the-art and the future. EMNLP (tutorial)

Crook, B., Schlüter, M., & Speith, T. (2023). Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). arXiv:2307.14239

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Aff, 92*, 28.

Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics, 9*, 1249–1267. https://doi.org/10.1162/tacl_a_00425, https://aclanthology.org/2021.tacl-1.74

De Cooman, J. (2022). Humpty dumpty and high-risk ai systems: The ratione materiae dimension of the proposal for an eu artificial intelligence act. *Mkt & Competition L Rev, 6*, 49.

De Hert, P. (2012). A human rights perspective on privacy and data protection impact assessments. *Privacy impact assessment* (pp. 33–76). Springer.

De Terwangne, C. (2020). *Principles relating to processing of personal data* (pp. 309–320). Oxford University Press.

De Terwangne, C. (2020). Principles relating to processing of personal data. *The EU general data protection (GDPR): a commentary* (pp. 309–320). Oxford University Press.

Delobelle, P., Tokpo, E., & Calders, T., et al. (2022). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 1693–1706https://doi.org/10.18653/v1/2022.naacl-main.122, https://aclanthology.org/2022.naacl-main.122

Demetzou, K. (2019). Data protection impact assessment: A tool for accountability and the unclarified concept of 'high risk' in the general data protection regulation. *Computer Law & Security Review, 35*(6), 105342. https://doi.org/10.1016/j.clsr.2019.105342, https://www.sciencedirect.com/science/article/pii/S0267364918304357

Demetzou, K. (2019). Data protection impact assessment: A tool for accountability and the unclarified concept of 'high risk' in the general data protection regulation. *Computer Law & Security Review, 35*(6), 105342.

Dev, S., Sheng, E., & Zhao, J., et al. (2022). On measures of biases and harms in NLP. In: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022. Association for Computational Linguistics, Online only, pp 246–267, https://aclanthology.org/2022.findings-aacl.24

Devlin, J., Chang, M.W., & Lee, K., et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805, arXiv:1810.04805

Diederich, J., Kindermann, J., Leopold, E., et al. (2003). Authorship attribution with support vector machines. *Applied Intelligence, 19*, 109–123.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way,*. Springer.

Ding, S.H.H., Fung, B.C.M., & Debbabi, M. (2015). A visualizable evidence-driven approach for authorship attribution. ACM Trans Inf Syst Secur 17(3https://doi.org/10.1145/2699910,

Dixon, L., Li, J., & Sorensen, J., et al. (2018). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, AIES '18, p 67–73, https://doi.org/10.1145/3278721.3278729,

Dubberley, S., Griffin, E., & Bal, H. M. (2015). *Making secondary trauma a primary issue: A study of eyewitness media and vicarious trauma on the digital frontline* (pp. 1–69). Eyewitness Media Hub.

Duran, F., & Woodhams, J. (2022). Impact of traumatic material on professionals in analytical and secondary investigative roles working in criminal justice settings: a qualitative approach. *Journal of Police and Criminal Psychology, 37*(4), 904–917.

Edwards, L. (2021). The eu ai act: a summary of its significance and scope. Artificial Intelligence (the EU AI Act) 1

Enriquez, D., Christensen, G., & Donovan, H., et al. (2023). Authorship verification for hired plagiarism detection. In: Proceedings of the 9th International Conference on Applied Computing & Information Technology. Association for Computing Machinery, New York, NY, USA, ACIT '22, p 19–24, https://doi.org/10.1145/3543895.3543928,

Escart'in, C.P., Lynn, T., & Moorkens, J., et al. (2021). Towards transparency in nlp shared tasks. ArXiv abs/2105.05020

Ethayarajh, K., & Jurafsky, D. (2021). Utility is in the eye of the user: A critique of nlp leaderboards. arXiv:2009.13888

Felzmann, H., Villaronga, E. F., Lutz, C., et al. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 2053951719860542.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., et al. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Fjeld, J., Achten, N., & Hilligoss, H., et al. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. Berkman Klein Center Research Publication (2020)

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review 1(1). Https://hdsr.mitpress.mit.edu/pub/l0jsh9d1

Floridi, L., et al. (2021). *Ethics, governance, and policies in artificial intelligence*. Springer.

Fobbe, E. (2021). Text-linguistic analysis in forensic authorship attribution

Frye, R.H., & Wilson, D.C. (2018). Defining forensic authorship attribution for limited samples from social media. In: The Thirty-First International Flairs Conference

Gasser, U. (2016). Recoding privacy law: Reflections on the future relationship among law, technology, and privacy. *Harv L Rev F, 130*, 61.

Gebru, T., Morgenstern, J., & Vecchione, B., et al. (2021). Datasheets for datasets. arXiv:1803.09010

Gellman, R. (2022). Fair information practices: A basic history-version 2.22. Available at SSRN

Gerards, J., Schäfer, M.T., & Muis, I., et al. (2022). Fundamental rights and algorithms impact assessment (fraia)

Gold, J. A., Zaremski, M. J., Lev, E. R., et al. (1993). Daubert v merrell dow: The supreme court tackles scientific evidence in the courtroom. *JAMA, 270*(24), 2964–2967.

Grünewald, E., & Pallas, F. (2021). Tilt: A gdpr-aligned transparency information language and toolkit for practical privacy engineering. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp 636–646

Gupta, A., Thadani, K., & O'Hare, N. (2020). Effective few-shot classification with transfer learning. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp 1061–1066, https://doi.org/10.18653/v1/2020.coling-main.92, https://aclanthology.org/2020.coling-main.92

Guthrie, D., Guthrie, L., & Allison, B., et al. (2007). Unsupervised anomaly detection. pp 1624–1628

Habernal, I., Mireshghallah, F., & Thaine, P., et al. (2023). Privacy-preserving natural language processing. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts. Association for Computational Linguistics, Dubrovnik, Croatia, pp 27–30, https://aclanthology.org/2023.eacl-tutorials.6

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. Common Market Law Review 55(4)

Hacker, P., & Passoth, J.H. (2020). Varieties of ai explanations under the law. from the gdpr to the aia, and beyond. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer, pp 343–373

Hacker, P., Cordes, J., & Rochon, J. (2022). Regulating gatekeeper ai and data: Transparency, access, and fairness under the dma, the gdpr, and beyond. arXiv preprint arXiv:2212.04997

Haduong, N., Gao, A., & Smith, N.A. (2023). Risks and NLP design: A case study on procedural document QA. In: Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp 1248–1269, https://aclanthology.org/2023.findings-acl.81

Halvani, O., & Graner, L. (2021). Posnoise: An effective countermeasure against topic biases in authorship analysis. In: Proceedings of the 16th International Conference on Availability, Reliability and Security. Association for Computing Machinery, New York, NY, USA, ARES 21, https://doi.org/10.1145/3465481.3470050,

Hämäläinen, M., & Alnajjar, K. (2021). The great misalignment problem in human evaluation of NLP methods. In: Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval). Association for Computational Linguistics, Online, pp 69–74, https://aclanthology.org/2021.humeval-1.8

Hessenthaler, M., Strubell, E., & Hovy, D., et al. (2022). Bridging fairness and environmental sustainability in natural language processing. arXiv:2211.04256

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 15*(8), e12432.

Hovy, D., & Spruit, S.L. (2016). The social impact of natural language processing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Berlin, Germany, pp 591–598, https://doi.org/10.18653/v1/P16-2096, https://aclanthology.org/P16-2096

Howard, B.S. (2008). Authorship attribution under the rules of evidence: empirical approaches–a layperson's legal system. International Journal of Speech, Language & the Law 15(2)

Huertas-Tato, J., Martín, A., & Huertas-García, Á., et al. (2022). Generating authorship embeddings with transformers. 2022 International Joint Conference on Neural Networks (IJCNN) pp 1–8. https://api.semanticscholar.org/CorpusID:252626603

Hupont, I., Micheli, M., Delipetrev, B., et al. (2023). Documenting high-risk ai: a european regulatory perspective. *Computer, 56*(5), 18–27.

Iqbal, F., Hadjidj, R., Fung, B. C., et al. (2008). A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation, 5*, S42–S51. https://doi.org/10.1016/j.diin.2008.05.001, https://www.sciencedirect.com/science/article/pii/S1742287608000315, the Proceedings of the Eighth Annual DFRWS Conference

Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices, 70*, 978–981.

Jafariakinabad, F., Tarnpradab, S., & Hua, K.A. (2019). Syntactic recurrent neural network for authorship attribution. arXiv preprint arXiv:1902.09723

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

John Albert AMSarah Michot, & Müller, A. (2022). Policy Brief: Our recommendations for strengthening data access for public interest research - AlgorithmWatch — algorithmwatch.org. https://algorithmwatch.org/en/policy-brief-platforms-data-access/, [Accessed 17-Jul-2023]

John-Mathews, J. M., Cardon, D., & Balagué, C. (2022). From reality to world a critical perspective on ai fairness. *Journal of Business Ethics, 178*(4), 945–95. https://doi.org/10.1007/s10551-022-05055-8

Julian, H., & van den Berg Esther, Ines R. (2017). Authorship attribution with convolutional neural networks and POS-eliding. In: Proceedings of the Workshop on Stylistic Variation. Association for Computational Linguistics, Copenhagen, Denmark, pp 53–58, https://doi.org/10.18653/v1/W17-4907, https://aclanthology.org/W17-4907

Juola, P. (2020). Authorship studies and the dark side of social media analytics. *Journal of Universal Computer Science, 26*, 156–170. https://doi.org/10.3897/jucs.2020.009

Kale, S. D., & Prasad, R. S. (2017). A systematic review on author identification methods. *Int J Rough Sets Data Anal, 4*, 81–91.

Keenan Jones SLJason, R. & Nurse, C. (2022). Are you robert or roberta? deceiving online authorship attribution models using neural text generators. arXiv:2203.09813

Kennedy, E., & Millard, C. (2016). Data security and multi-factor authentication: Analysis of requirements under eu law and in selected eu member states. *Computer Law & Security Review, 32*(1), 91–110. https://doi.org/10.1016/j.clsr.2015.12.004, https://www.sciencedirect.com/science/article/pii/S0267364915001697

Khonji, M., Iraqi, Y., & Jones, A. (2015). An evaluation of authorship attribution using random forests. In: 2015 international conference on information and communication technology research (ictrc), IEEE, pp 68–71

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in ai decision support: Explaining and visualizing convolutional neural networks for text information. *Decis Support Syst, 134*, 113302.

Kirk, H., Birhane, A., & Vidgen, B., et al. (2022). Handling and presenting harmful text in NLP research. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 497–510, https://aclanthology.org/2022.findings-emnlp.35

Klymenko, O., Meisenbacher, S., & Matthes, F. (2022). Differential privacy in natural language processing the story so far. In: Proceedings of the Fourth Workshop on Privacy in Natural Language Processing. Association for Computational Linguistics, Seattle, United States, pp 1–11, https://doi.org/10.18653/v1/2022.privatenlp-1.1, https://aclanthology.org/2022.privatenlp-1.1

Kondyurin, I. (2022). Explainability of transformers for authorship attribution. Master's thesis

Koops, B. J. (2021). The concept of function creep. *Law, Innovation and Technology, 13*, 1–28. https://doi.org/10.1080/17579961.2021.1898299

Koppel, M., & Schler, J. (2004). Authorship verification as a one-class classification problem. Proceedings of the twenty-first international conference on Machine learning

Koppel, M., Argamon, S.E., & Shimoni, A.R. (2002). Automatically categorizing written texts by author gender. Lit Linguistic Comput 17:401–412. https://api.semanticscholar.org/CorpusID:1057413

Kumar, R., Yadav, S., & Daniulaityte, R., et al. (2020). Edarkfind: Unsupervised multi-view learning for sybil account detection. In: Proceedings of The Web Conference 2020. Association for Computing Machinery, New York, NY, USA, WWW '20, p 1955–1965, https://doi.org/10.1145/3366423.3380263,

Lacoste, A., Luccioni, A., & Schmidt, V., et al. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700

Lalor, J., Yang, Y., & Smith, K., et al. (2022). Benchmarking intersectional biases in NLP. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 3598–3609, https://doi.org/10.18653/v1/2022.naacl-main.263, https://aclanthology.org/2022.naacl-main.263

Laufer, B., Jain, S., & Cooper, A.F., et al. (2022). Four years of facct: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22, p 401–426, https://doi.org/10.1145/3531146.3533107,

Lawrence, S., & Giles, C.L. (2000). Overfitting and neural networks: conjugate gradient and backpropagation. In: Proceedings of the

IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, IEEE, pp 114–119

Lei, Z., Qi, H., & Han, Y., et al. (2022). Application of bert in author verification task. In: Conference and Labs of the Evaluation Forum

Locatelli, M., Tagliabue, L.C., & Di Giuda, G.M., et al. (2022). Archiberto: a hierarchization quality objectives nlp tool in the italian architecture, engineering and construction sector. In: CEUR WORKSHOP PROCEEDINGS, Lops, Pasquale; Basile, Pierpaolo; Siciliani, Lucia; Taccardi, Vincenzo; Di ..., pp 8–25

Loi, M., & Spielkamp, M. (2021). Towards accountability in the use of artificial intelligence for public administrations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp 757–766

Loi, M., Mätzener, A., Müller, A., et al. (2021). *Automated decision-making systems in the public sector an impact assessment tool for public authorities*. AW AlgorithmWatch gGmbH.

Lund, B. D., Wang, T., Mannuru, N. R., et al. (2023). chatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology, 74*(5), 570–581. https://doi.org/10.1002/asi.24750

Lynskey, O. (2017). The 'europeanisation' of data protection law. *Cambridge Yearbook of European Legal Studies, 19*, 252–286.

Ma, P., Wang, S., & Liu, J. (2020). Metamorphic testing and certified mitigation of fairness violations in nlp models. In: International Joint Conference on Artificial Intelligence

Madiega, T. (2021). *Artificial intelligence act*. European Parliament: European Parliamentary Research Service.

Manolache, A., Brad, F., & Burceanu, E., et al. (2021). Transferring bert-like transformers' knowledge for authorship verification. arXiv preprint arXiv:2112.05125

Manolache, A., Brad, F., & Barbalau, A., et al. (2022). Veridark: A large-scale benchmark for authorship verification on the dark web. arXiv:2207.03477

Mantelero, A. (2018). Ai and big data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review, 34*(4), 754–772.

Martin, N., Friedewald, M., & Schiering, I., et al. (2020a). The Data Protection Impact Assessment According to Article 35 GDPR. Fraunhofer Verlag, https://doi.org/10.24406/publica-fhg-300244, https://publica.fraunhofer.de/handle/publica/300244

Martin, N., Friedewald, M., & Schiering, I., et al. (2020b). The data protection impact assessment according to article 35 gdpr

Mateless, R., Tsur, O., & Moskovitch, R. (2021). Pkg2vec: Hierarchical package embedding for code authorship attribution. *Future Generation Computer Systems, 116*, 49–60.

Mehrabi, N., Morstatter, F., & Saxena, N., et al. (2022). A survey on bias and fairness in machine learning. arXiv:1908.09635

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence, 1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mohsen, A.M., El-Makky, N.M., & Ghanem, N.M. (2016). Author identification using deep learning. 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) pp 898–903

Mollen, A. (2023). New study highlights crucial role of trade unions for algorithmic transparency and accountability in the world of work - AlgorithmWatch — algorithmwatch.org. https://algorithmwatch.org/en/study-trade-unions-algorithmic-transparency/, [Accessed 17-Jul-2023]

Mondschein CF, Monda C (2019) The EU's General Data Protection Regulation (GDPR) in aÂ Research Context, Springer International Publishing, Cham, pp 55–71. https://doi.org/10.1007/978-3-319-99713-1_5,

Moraes, T. G., Almeida, E. C., & de Pereira, J. R. L. (2021). Smile, you are being identified! risks and measures for the use of facial recognition in (semi-)public spaces. *AI and Ethics, 1*(2), 159–172. https://doi.org/10.1007/s43681-020-00014-3

Murauer, B., & Specht, G. (2021a). Developing a benchmark for reducing data bias in authorship attribution. In: Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 179–188, https://doi.org/10.18653/v1/2021.eval4nlp-1.18, https://aclanthology.org/2021.eval4nlp-1.18

Murauer, B., & Specht, G. (2021b). Developing a benchmark for reducing data bias in authorship attribution. In: Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, pp 179–188

Neme, A., Lugo, B., & Cervera, A. (2011). Authorship attribution as a case of anomaly detection: A neural network model. *Int J Hybrid Intell Syst, 8*, 225–235.

Neuwirth, R. J. (2023). Prohibited artificial intelligence practices in the proposed eu artificial intelligence act (aia). *Computer Law & Security Review, 48*, 105798.

Nirkhi, S., & Dharaskar, Dr.R.V. (2013). Comparative study of authorship identification techniques for cyber forensics analysis. International Journal of Advanced Computer Science and Applications 4(5). https://doi.org/10.14569/IJACSA.2013.040505,

Nissim, K., & Wood, A. (2018). Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170358.

OECD (2013) OECD Legal Instruments — legalinstruments.oecd.org. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0188, [Accessed 20-Jul-2023]

Panov, V., Kovalchuk, M., Filatova, A., et al. (2022). Mucaat: Multilingual contextualized authorship anonymization of texts from social networks. *Procedia Computer Science, 212*, 322–329. https://doi.org/10.1016/j.procs.2022.11.016, https://www.sciencedirect.com/science/article/pii/S1877050922017070, 11th International Young Scientist Conference on Computational Science

Plank, B., Hovy, D., & McDonald, R., et al. (2014). Adapting taggers to Twitter with not-so-distant supervision. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp 1783–1792, https://aclanthology.org/C14-1168

Potthast, M., Hagen, M., Stein, B. (2016). Author obfuscation: Attacking the state of the art in authorship verification. In: Conference and Labs of the Evaluation Forum

Powers, D.M. (2008). Evaluation evaluation. In: ECAI 2008. IOS Press, p 843–844

Prabhu, A., Dognin, C., & Singh, M. (2019). Sampling bias in deep active classification: An empirical study. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 4058–4068, https://doi.org/10.18653/v1/D19-1417, https://aclanthology.org/D19-1417

Prasad, S.N., Narsimha, V., & Reddy, P.V., et al. (2015). Influence of lexical, syntactic and structural features and their combination on authorship attribution for telugu text. Procedia Computer Science 48:58–64. https://doi.org/10.1016/j.procs.2015.04.110, https://www.sciencedirect.com/science/article/pii/S1877050915006195, international Conference on Computer, Communication and Convergence (ICCC 2015)

Prem, E. (2023). From ethical ai frameworks to tools: a review of approaches. *AI and Ethics, 3*(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Procter, R.N., Rouncefield, M., & Tolmie, P. (2020). Accounts, accountability and agency for safe and ethical ai. ArXiv abs/2010.01316

Pyevich, C. M., Newman, E., & Daleiden, E. (2003). The relationship among cognitive schemas, job-related traumatic exposure, and posttraumatic stress disorder in journalists. *Journal of Traumatic Stress: Official Publication of the International Society for Traumatic Stress Studies, 16*(4), 325–328.

Qian, K., Danilevsky, M., & Katsis, Y., et al. (2021). Xnlp: A living survey for xai research in natural language processing. In: 26th International Conference on Intelligent User Interfaces - Companion. Association for Computing Machinery, New York, NY, USA, IUI '21 Companion, p 78–80, https://doi.org/10.1145/3397482.3450728,

Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Rawal, A., McCoy, J., & Rawat, D.B., et al. (2021). Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives. IEEE Transactions on Artificial Intelligence PP:1–1

Regulation, P. (2018). *General data protection regulation. Intouch, 25*, 1–5.

Renda, A., Arroyo, J., Fanni, R., et al. (2021). *Study to support an impact assessment of regulatory requirements for artificial intelligence in europe*. Brussels.

Rudin, C.(2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. https://arxiv.org/abs/1811.10154, arXiv:1811.10154

Ruparelia, N. B. (2010). Software development lifecycle models. *SIGSOFT Softw Eng Notes, 35*(3), 8–13. https://doi.org/10.1145/1764810.1764814

Saeed, W., & Omlin, C. (2023). Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems, 263*, 110273. https://doi.org/10.1016/j.knosys.2023.110273, https://www.sciencedirect.com/science/article/pii/S0950705123000230

Salur, M. U., & Aydin, I. (2020). A novel hybrid deep learning model for sentiment classification. *IEEE Access, 8*, 58080–58093.

Sapkota, U., Bethard, S., & Montes, M., et al. (2015). Not all character n-grams are created equal: A study in authorship attribution. In: Mihalcea R, Chai J, Sarkar A (eds) Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Denver, Colorado, pp 93–102, https://doi.org/10.3115/v1/N15-1010, https://aclanthology.org/N15-1010

Saxena, V., Bashpole, B., & van Dijck, G., et al. (2023a). Idtraffickers: An authorship attribution dataset to link and connect potential human-trafficking operations on text escort advertisements. arXiv:2310.05484

Saxena, V., Rethmeier, N., & van Dijck, G., et al. (2023b). VendorLink: An NLP approach for identifying & linking vendor migrants & potential aliases on Darknet markets. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 8619–8639, https://aclanthology.org/2023.acl-long.481

Saxon, M.S., Levy, S., & Wang, X., et al. (2021). Modeling disclosive transparency in nlp application descriptions. In: Conference on Empirical Methods in Natural Language Processing

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning, 10*, 153–178.

Sennewald, B., Herpers, R., & Hülsmann, M., et al. (2020). Voting for authorship attribution applied to dark web data. In: Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering, pp 217–226

Shah, D., Schwartz, H.A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. arXiv preprint arXiv:1912.11078

Shah, D.S., Schwartz, H.A., & Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 5248–5264, https://doi.org/10.18653/v1/2020.acl-main.468, https://aclanthology.org/2020.acl-main.468

Shamsi, J. A., Zeadally, S., Sheikh, F., et al. (2016). Attribution in cyberspace: Techniques and legal implications. *Security and Communication Networks, 9*(15), 2886–2900.

Shmueli, B., Fell, J., & Ray, S., et al. (2021). Beyond fair pay: Ethical implications of nlp crowdsourcing. arXiv:2104.10097

Shook, J., Smith, R., & Antonio, A. (2017). Transparency and fairness in machine learning applications. *Tex A &M J Prop L, 4*, 443.

Shrestha, P., Sierra, S., & González, F., et al. (2017). Convolutional neural networks for authorship attribution of short texts. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valencia, Spain, pp 669–674, https://aclanthology.org/E17-2106

Silva, K., Can, B., & Blain, F., et al. (2023). Authorship attribution of late 19th century novels using GAN-BERT. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). Association for Computational Linguistics, Toronto, Canada, pp 310–320, https://doi.org/10.18653/v1/2023.acl-srw.44, https://aclanthology.org/2023.acl-srw.44

Simbeck, K. (2022). Facct-check on ai regulation: Systematic evaluation of ai regulation on the example of the legislation on the use of ai in the public sector in the german federal state of schleswig-holstein. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22, p 89–96, https://doi.org/10.1145/3531146.3533076,

Sion, L., Van Landuyt, D., & Joosen, W. (2021). An overview of runtime data protection enforcement approaches. In: 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS &PW), IEEE, pp 351–358

Sjöberg, C.M. (2021). Legal ai from a privacy point of view: Data protection and transparency in focus. Digital Human Sciences p 181

Søgaard, A., Plank, B., & Hovy, D. (2014). Selection bias, label bias, and bias in ground truth. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts, pp 11–13

Solanke, A. A. (2022). Explainable digital forensics ai: Towards mitigating distrust in ai-based digital forensics analysis using interpretable models. *Forensic Science International: Digital Investigation, 42*, 301403. https://doi.org/10.1016/j.fsidi.2022.301403, https://www.sciencedirect.com/science/article/pii/S2666281722000841, proceedings of the Twenty-Second Annual DFRWS USA

Solove, D. J. (2005). A taxonomy of privacy. *U Pa l Rev, 154*, 477.

Sousa, S., & Kern, R. (2023). How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review, 56*(2), 1427–1492. https://doi.org/10.1007/s10462-022-10204-6

Srinivasan, R., & Chander, A. (2021). Biases in ai systems: A survey for practitioners. *Queue, 19*(2), 45–64. https://doi.org/10.1145/3466132.3466134

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538–556. https://doi.org/10.1002/asi.21001, https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001

Stamatatos, E., et al. (2006). Ensemble-based author identification using character n-grams. In: Proceedings of the 3rd International Workshop on Text-based Information Retrieval, pp 41–46

Staunton, C., Slokenberga, S., & Mascalzoni, D. (2019). The gdpr and the research exemption: Considerations on the necessary safeguards for research biobanks. *European Journal of Human Genetics, 27*(8), 1159–1167.

Stepin, I., Alonso, J. M., Catala, A., et al. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access, 9*, 11974–12001.

Stevens, A., Deruyck, P., & Veldhoven, Z.V., et al. (2020). Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp 1241–1248, https://doi.org/10.1109/SSCI47803.2020.9308371

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. https://arxiv.org/abs/1906.02243, arXiv:1906.02243

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization. ACM, https://doi.org/10.1145/3465416.3483305,

Sweeney, L., Crosas, M., & Bar-Sinai, M. (2015). Sharing sensitive data with confidence: The datatags system. Technology Science

Tabassi, E. (2023). Artificial intelligence risk management framework (ai rmf 1.0)

Tamo-Larrieux, A. (2021). Decision-making by machines: Is the 'law of everything' enough? *Computer Law & Security Review, 41*, 105541.

Tamò-Larrieux, A. (2018). Designing for Privacy and its Legal Framework — link.springer.com. https://link.springer.com/book/10.1007/978-3-319-98624-1, [Accessed 18-Jul-2023]

Theophilo, A., Padilha, R., & Andaló, F.A., et al. (2022). Explainable artificial intelligence for authorship attribution on social media. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2909–2913, https://doi.org/10.1109/ICASSP43922.2022.9746262

Tubella, A.A., Theodorou, A., & Dignum, V., et al. (2019). Governance by glass-box: Implementing transparent moral bounds for ai behaviour. arXiv preprint arXiv:1905.04994

Uchendu, A., Le, T., & Lee, D. (2023). Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 25*(1), 1–18.

Van Es, K., & Schäfer, M. T. (2017). *The datafied society*. Amsterdam University Press.

Van Wynsberghe, A. (2021). Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics, 1*(3), 213–218.

Voigt, P., & Von dem Bussche, A. (2017). *The eu general data protection regulation (gdpr). A practical guide* (1st ed., pp. 10–5555). Springer International Publishing.

Voigt, P., & Avd, Bussche. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide* (1st ed.). Springer Publishing Company.

Voisin, G., Boardman, R., & Assion, S., et al.(2020). Ico, cnil, german and spanish dpa revised cookies guidelines: Convergence and divergence. Recuperado de https://iapporg/media/pdf/resource_center/CNIL_ICO_chartpdf

Wegmann, A., Schraagen, M., & Nguyen, D. (2022). Same author or just same topic? towards content-independent style representations. In: Proceedings of the 7th Workshop on Representation Learning for NLP. Association for Computational Linguistics, Dublin, Ireland, pp 249–268, https://doi.org/10.18653/v1/2022.repl4nlp-1.26, https://aclanthology.org/2022.repl4nlp-1.26

Weidinger, L., Mellor, J., & Rauh, M., et al. (2021). Ethical and social risks of harm from language models. arXiv:2112.04359

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ml fairness approaches. *Journal of Artificial Intelligence Research, 74*, 75–109.

Wolfe, R., & Caliskan, A. (2021). Low frequency names exhibit bias and overfitting in contextualizing language models. arXiv preprint arXiv:2110.00672

Wright, D., & May, A. (2014). Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law (Linguagem e Direito), 1*, 37–69.

Wu, C. J., Raghavendra, R., Gupta, U., et al. (2022). Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems, 4*, 795–813.

Yang, M., & Chow, K. P., et al. (2014). Authorship attribution for forensic investigation with thousands of authors. In N. Cuppens-Boulahia, F. Cuppens, & S. Jajodia (Eds.), *ICT Systems Security and Privacy Protection* (pp. 339–350). Berlin Heidelberg, Berlin, Heidelberg: Springer.

Yang, M., & Chow, K.P. (2014b). Authorship attribution for forensic investigation with thousands of authors. In: ICT Systems Security and Privacy Protection: 29th IFIP TC 11 International Conference, SEC 2014, Marrakech, Morocco, June 2-4, 2014. Proceedings 29, Springer, pp 339–350

Yenduri, G., M R, G CS, & et al. (2023). Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv:2305.10435

Young, M., Katell, M., & Krafft, P. (2022). Confronting power and corporate capture at the facct conference. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22, p 1375–1386, https://doi.org/10.1145/3531146.3533194,

Yuluce, I., & Dalkıç, F. (2022). Author identification with machine learning algorithms. International Journal of Multidisciplinary Studies and Innovative Technologies 6:45. https://doi.org/10.36287/ijmsit.6.1.45

Zafar, M.B., Valera, I., & Rodriguez, M.G., et al. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, https://doi.org/10.1145/3038912.3052660,

Zhai, W., Rusert, J., & Shafiq, Z., et al. (2022). Adversarial authorship attribution for deobfuscation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 7372–7384, https://doi.org/10.18653/v1/2022.acl-long.509, https://aclanthology.org/2022.acl-long.509

Zhang, J., Shu, Y., & Yu, H. (2023). Fairness in design: A framework for facilitating ethical artificial intelligence designs. *International Journal of Crowd Science, 7*(1), 32–39. https://doi.org/10.26599/IJCS.2022.9100033

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., et al. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST), 11*(3), 1–41.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in neural information processing systems 28

Zhang, Y., Fan, Y., & Song, W., et al. (2019). Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In: The World Wide Web Conference. Association for Computing Machinery, New York, NY, USA, WWW '19, p 3448–3454, https://doi.org/10.1145/3308558.3313537,

Zhang, Z., Strubell, E., & Hovy, E. (2022). A survey of active learning for natural language processing. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 6166–6190, https://aclanthology.org/2022.emnlp-main.414

Zheng, R., Qin, Y., Huang, Z., et al. (2003). Authorship analysis in cybercrime investigation. In H. Chen, R. Miranda, D. D. Zeng, et al. (Eds.), *Intelligence and Security Informatics* (pp. 59–73). Springer.