SciencesPo
SCHOOL OF PUBLIC AFFAIRS

PUBLIC POLICY MASTER THESIS

April 2022

Unregulated Negative Impacts of AI:

Mixed Methods Analysis of Feedback Responses to the EU AI Act Proposal

Martyna Kalvaitytė

Master's Thesis supervised by Jen Schradie

Second member of the Jury: Bertrand Pailhès

Master in Public Policy
Digital, New Technology and Public Policy

Abstract

Negative AI impacts are increasingly more noticeable, presenting regulators with the challenge of balancing the opportunities and risks associated with AI. The AI Act Proposal of the European Commission undertakes this challenging task. Two hundred sixty-six feedback responses to the Proposal are analysed using a proposed mixed method to tackle the question of what are the main negative impacts of AI that regulators have failed to address. The study contributes to the literature on adverse AI impacts by offering a mapping of the cross-sectoral impacts and noting their different qualities. Through topic modelling, it is found that the main negative AI impacts are centred around manipulation, the use of biometric recognition systems, adverse effects on workers and children's groups, and overarching potential human rights violations. Guided close reading of identified impact groups' most representative feedback responses illustrates that impacts are both individual and social, emphasising the issue of the lack of protections against societal level impacts. Close reading also provides a use case of algorithmic impacts' descriptions, exemplifying qualities of negative AI impacts outlined by Smuha (2021a) and Tufekci (2015). It is recommended to address the identified individual and social effects by creating protections against societal impacts and establishing redress mechanisms for claiming individual, communal and social remedies. Following identified agreement across investigated responses, it is recommended to establish an independent institution tasked with measuring and monitoring AI systems to increase the knowledge base surrounding the extent of negative AI impacts and the mechanics in which they come into being.

Key words
Artificial intelligence, AI governance, regulation, European Union, societal AI impacts, manipulation

Table of Contents

Acknowledgments

Thank you, Dr. Jen Schradie, for guiding me throughout this journey, for your advice on broadening the scope of research, for great recommendations, which got me out of the risky frame situation, and for offering critical suggestions at critical times. And thank you for drawing my attention to digital inequalities and how they manifest through the social aspects of technologies.

Thank you to three interviewees for openly sharing their understandings of social AI impacts and engaging in discussions at the conception stages of the research. The variety of approaches significantly shaped the decision to focus on social.

I would also like to thank Charleyne Biondi for introducing me to the surveillance studies field back at Sciences Po Paris Reims Campus. I would not have had so many formative professional experiences in the privacy field if it were not for the course, which ultimately led to AI, just in time for the field's most exciting of times, grappling with the accumulation of unaddressed issues.

With explicit consent from my friends for their names' disclosure, I would like to thank Chantal, David, Guillaume, Gytis, Jorūnė, Justas, and Viktorija. Thank you for keeping me grounded throughout the process; thank you for our discussions, your comments, editing and all the emotional support provided.

And ultimate thank you to my parents for their unconditional support.

Why should I read this research?

Technological AI advancements and their increased deployments present regulators with the challenge of balancing the benefits associated with AI technologies and mitigating negative risks and impacts. The European Commission's proposed AI Act (AIA) is an attempt to draw this balance. The European Parliament Committees are currently preparing their opinions on the Proposal, which will undergo its reading stages. Despite the AIA being a moving target, the analysis of unaddressed negative AI impacts provides a reference point to regulators and policymakers, which can be referred to throughout the legislative process.

Mixed methods of the computational topic modelling and guided close reading were applied in the analysis of 266 feedback responses to the European Commission's consultation on the AIA, addressing the research question of what the main negative AI impacts are unaddressed by regulation. The topics related to the negative effects are identified by analysing the feedback responses with a type of textual analysis, topic modelling. The select topics are further explored with guided close reading, mapping negative AI impacts and allowing the study of adverse effects' qualities and certainty levels of these effects as described by consultation's respondents.

At the time of writing this research, this study offers the first analysis of the feedback responses to the Proposal, apart from the descriptive statistics published by the European Commission (European Commission 2021b). This research contributes to the literature on algorithmic impacts by outlining the main negative AI impacts requiring timely resolutions, as noted by respondents to the European Commission's call for feedback. The study also provides a cross-sectoral mapping of negative algorithmic effects contributing to studies seeking to provide more consolidated and overarching approaches to AI impacts' mapping. A methodological contribution is also made to the policy field's uses of topic modelling by complementing the topic modelling technique with Nelson's (2020) guided close reading as embedded in computational grounded theory (Baumer et al. 2017; Nelson 2020).

Manipulation of users, adverse effects associated with biometric identification technologies, effects on workers and children's groups, and overarching human rights infringements identified across the effects mentioned above are identified as the main unaddressed negative AI impacts. The mapping of issue areas through close reading reveals that the adverse AI impacts transcend the individual level. Societal impacts affect users individually and on the societal level. These impacts present risks to deepening inequality due to discriminatory biometric recognition systems, societal polarisation, and increased misinformation caused by manipulation. The thesis echoes the arguments for the emergence of a new type of algorithmic impact, namely social level impacts (Smuha 2021a; Tufekci 2015), demonstrating a unique set of qualities. The emergence of social level impacts is crucial due to the absence of protections against these negative social impacts (Smuha 2021a). The widespread agreement across analysed documents signals the need to better understand and measure developing AI impacts.

The recommendations to tackle the negative AI impacts identified by this research fall into two main categories, one addressing the impacts and the other tackling the lack of knowledge surrounding them. The first category can be addressed on two fronts: either preventing the impacts from taking place or providing users with adequate mechanisms for seeking remedies after being exposed to negative individual or social AI impacts. The creation of an independent institution tasked with measuring and monitoring AI impacts is suggested to address the limited knowledge base surrounding AI effects.

1. Introduction

As AI[1] grows more prevalent (Chui, Hall, and Sukharevsky 2021), the number of cases of negative algorithmic impacts increases, exposing the need for regulatory intervention to protect digital users. Some recent examples include discrimination against women and other disadvantaged groups across human resources algorithms (Yam and Skorburg 2021) and discriminatory algorithms managing access to healthcare (Obermeyer et al. 2019). The public sphere is also affected by algorithms promoting political animosity on social media platforms (Rathje, Bavel, and Linden 2021). Given the constantly evolving field of algorithmic impacts, the AIAAIC Repository provides a documentation list of public algorithmic incidents, which as of 24 April, has 868 cases (AIAAIC n.d.). The types and range of spheres where impacts are exerted are as broad as the application areas of AI.

Negative algorithmic impacts are a result of the increasing uptake of AI technologies across everyday technologies, which by interacting with existing structures, produce changes (Maas 2021). These impacts receive increasing public (Crépel and Cardon 2021) and scholarly attention (Ganesh and Moss 2022). The changes brought by information technologies of which AI is a subset can be described as bringing the fourth (industrial) revolution (Floridi 2014; Schwab 2016). The study of negative AI impacts and the efficiency of proposed interventions to address them is of utmost importance given the speed of AI advancements, the rate of their application within widely used technologies (Crawford 2021), and the revolutionary scale of changes brought by AI. As a result of these high stakes, the focus of this thesis is placed on negative AI impacts that are not addressed by forthcoming AI regulation on a European level.

The feedback responses to the European Commission's (EC) Proposal for AI regulation of April 2021, also known as the AI Act (AIA) offer an exciting area for the study of statements on unaddressed negative AI impacts. The following questions guide the analysis of feedback responses: what are the main individual and societal negative AI impacts? How is the uncertain nature of algorithmic impacts reflected in feedback responses outlining the loopholes of the AIA?

These questions are tackled by the main research question: **What are the main negative impacts of AI that regulators have failed to address?** Reasons for focusing on the European Union (EU) and its AIA are explained below, together with a rationale for investigating impacts reaching beyond individual level effects.

Placing special attention on AI governance in the EU is imperative given its recent global technology regulatory dominance status, primarily established by the GDPR (also known as the Brussels effect) (Bradford 2020; Hadjiyianni 2020). The EU regulatory action is analysed instead of AI governance approaches taken by China or the US because China takes a more active approach to AI governance in designing for predetermined societal outcomes (Roberts et al. 2021). In the meantime, the US relies on self-regulation embedded in a libertarian approach to technologies (Feijóo et al. 2020). The EU is thus a significant field of investigation because it balances self-regulation with governmental regulation in its attempts to benefit from AI advances while mitigating arising risks.

[1] Definitions of AI and negative impacts are provided below in the 1.4. section on definitions.

The balance that the EU aims to strike is particularly important for managing AI impacts since self-regulation can result in negative societal impacts (Cusumano, Gawer, and Yoffie 2021). The efficiency of self-regulation depends on the perceived costs of regulation on the side of enterprises and the likelihood of intrusive governmental regulation (Cusumano et al. 2021:1279). Suppose the costs of implementing self-regulation are high while the likelihood of invasive governmental regulation is low. In that case, self-regulation's efficiency is lower, which can result in the creation of negative societal effects (Cusumano et al. 2021). Regarding the costs of addressing impacts created using AI, Zuboff (2019) explains current market relations and their use of personal data as surveillance capitalism, where wealth and power are created by using data to predict and shape human behaviour (Zuboff 2019). Despite one's evaluation of current market dynamics, the costs of self-regulation for technology companies are extremely high as the use of data is the main factor for their economic models. This balance aimed by the EU at leveraging self-regulation and governmental regulation explains the tension in managing the benefits and risks brought by AI. The process of balancing risks and benefits thus deserves a focus on negative AI impacts, such as manipulation and human rights infringements, because they can arise in regulatory contexts relying on self-regulation. The negative impacts can be overlooked given the strategic importance of advancing AI and taking a dominant market position (Smuha 2021b), which highlights the stakes of this study.

The EC Proposal for AI regulation represents an attempt at a balanced approach. The Explanatory Memorandum states that the Act aims "to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market" (European Commission 2021c). The AIA is part of the EC's strategy for AI, which includes the liability regulations package, the Digital Services Act (DSA), competition regulation, the Digital Markets Act (DMA), along with updated sectoral laws for product safety (European Commission 2021a). The AIA sets rules for the creation and use of AI systems that are classified according to levels of risk. The levels of risks determine different requirements or options for voluntary actions. According to the AIA Title II, unacceptable risks according to which certain practices are prohibited include manipulative systems, social scoring, and biometric systems for certain law enforcement uses. The prohibition of such uses can be regarded as failing to limit those uses (Veale and Borgesius 2021:100). The requirements are set for high-risk AI systems, as outlined in Annex III of the AIA, covering AI system areas, such as biometric identification, education, employment, and public services, among others. High-risk systems are required to meet data quality criteria, establish risk management systems, meet obligations to ensure accuracy, cybersecurity, and robustness of their systems, as well as carry out conformity assessments for which standards are to be created (Veale and Borgesius 2021:103). The high-risk applications' requirements are subject to self-regulation, with only biometric applications requiring an external body to conduct the assessment (Veale and Borgesius 2021:106). As outlined by Cusumano et al. (2021), self-regulation can lead to the production of negative impacts. The low-risk systems can voluntarily choose to meet the requirements set for high-risk systems. The AIA also creates a centralised public database containing instructions for the high-risk systems concerned. The draft AIA received criticism along the spectrum from corporations, other European Institutions (EDPB-EDPS 2021) and civil society for failing to create mechanisms for individual or community appeals (Veale and Borgesius 2021).

The revolutionary nature of changes brought by AI leads to the emergence of new types of impacts, some of which are not covered by existing protections. The impacts that go beyond the individual level include communal and societal levels of impact. These impacts warrant special attention since the approaches to protections against individual negative impacts take precedence, while social impacts tend to be overlooked (Campolo et al. 2017; Griffin 2022; Smuha 2021a). The overlap between individual and societal can take place; however, societal harms have a different scope of impact by affecting societal interests (Smuha 2021a:5). Societal harms warrant attention because they are more difficult to establish since they tend to be caused by repeated interactions that lead to incremental changes, which result in noticeable impacts after longer periods of time than individual harms (Smuha 2021a). The challenges of addressing societal impacts can also be interpreted through the direction of present AI governance approaches as problem-solving instead of problems-finding (Liu and Maas 2021). In the problem-solving governance mode, issues that have already been identified can receive more attention than problems that are new, such as societal AI impacts that challenge boundaries set by existing legal systems (Liu and Maas 2021). As a result, the thesis places special attention on social impacts, given the high stakes of negative impacts against which protections do not exist.

1.1. Research Question, Hypotheses and Method

The main research question of the thesis is **what are the main negative impacts of AI that regulators have failed to address**? A mixed methods approach is used to find topics generated by applying a quantitative data analysis technique, topic modelling (TM), of the feedback responses to the AIA. Close reading focusing on contextualising themes discovered by topic modelling is used as part of the sequential design in the analysis of a selected subset of themes evolving around negative AI impacts. Close reading is inspired by grounded theory (Baumer et al. 2017; Nelson 2020), meaning that additional subcategories to the main themes identified by TM are included in the analysis.

The following hypotheses are examined:

H1. Negative AI impacts are not only individual.

H2. The extent of negative AI impacts is uncertain.

The analysis of feedback responses to the AIA allows for investigating the wide scope of impacts ranging across different sectors, application areas, and levels of influence while at the same time allowing for the discussion of impacts in greater detail. Feedback respondents could upload one document and fill the text field for response; this format allowed respondents to provide more detailed statements regarding the AIA.

The themes of the AIA feedback responses focusing on negative AI impacts are identified through textual analysis, namely TM. From 24 topics generated, 5 topics are chosen for close reading analysis. These five topics are: the use of biometric identification technologies and associated impacts, manipulation, effects on groups such as workers and children, and human rights infringements. These topics are selected for the further analysis conducted with the close reading of the top 10 documents with the highest weights of selected topics produced by the TM model.

Inspired by Prunkl and Whittlestone's (2020) research on bridging the gap between near-term and long-term AI impacts, their suggested dimensions for situating research on AI, such as AI capabilities, impacts and certainty levels, are used for close reading analysis and adapted to the context of the AIA feedback analysis. The feedback responses' discourse around AI technologies and the changes they brought are analysed through the lens of technological capabilities. This analysis is guided by questions such as what AI techniques are discussed or are AI technologies discussed in a more general way, falling under the AI umbrella term. The questions of how and who is impacted, and what types of impact(s) are outlined aim to capture effects identified by feedback respondents as failed to be addressed by the AIA. The level of certainty is investigated by inspecting the questions of whether the levels of certainty on algorithmic impacts are expressed and how AI impacts are presented. The analysis via the lens of these categories is thus topic- and response-dependent.

H1 is tested by analysing negative impacts outlined by the mixed methods approach of TM and guided close reading. TM is used to inspect whether topics revolving around individual and societal impacts are present, while the interpretation of topics is then verified with guided close reading (Nelson 2020). Themes identified by TM are significant enough to be discussed in feedback responses; the responses are crafted in agenda-setting contexts shaped by limited resources and attention (Jones and Baumgartner 2005). The recurrence of themes establishes their importance since TM captures these topics as distinct from remaining topics across the entire corpora, representing all documents analysed with the TM model. The generated topics are distinct and grouped into their groups of related words. As a result, a generated topic can be interpreted as significant enough across the feedback responses to have its own topic.

Guided close reading is approached from grounded theory perspective (Nelson 2020); the topics identified by the TM model are adjusted according to the themes discussed by the feedback responses most representative of the topics. Thus, additional negative impacts are attributed to the themes investigated with close reading. Close reading is used to contextualise the topics revolving around individual and societal AI impacts. The resulting impacts are categorised into individual and social as interpreted by the respondents.

H2 is tested by focusing the lens of close reading on the expressed levels of uncertainty in the feedback responses or their absence. By nature, policymaking occurs in a place of uncertainty, which is a quality associated with innovation (Stilgoe, Owen, and Macnaghten 2013) and future-oriented processes (van Dorsser et al. 2018). Mitigation of negative AI impacts is particularly challenging due to the existing informational asymmetries on algorithmic impacts and the absence of governmental measuring and monitoring of incremental changes bringing societal AI impacts (Smuha 2021a). The level of uncertainty is investigated across discussions of negative AI impacts on selected topics. The uncertainty is approached in terms of respondents' statements of certainty regarding the impacts, noting the evidence base to support their impacts and through respondents' suggestions for increased monitoring of AI impacts.

1.2. Definitions

There is no clear consensus on the definition of the term AI (Krafft et al. 2020; Wang 2019). The definition set by the EC in the AIA is used in the thesis, given that it sets the context for feedback responses. According to Article 3(1):

> (1)'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. (European Commission 2021c)

The technologies laid down in Annex I include different forms of machine learning (ML), logic, knowledge-based and statistical approaches. For this thesis, AI impacts and algorithmic impacts will be used interchangeably, given the current status of the prevalence of supervised learning algorithms (Loukides 2021; Stanford Institute for Human-Centered Artificial Intelligence 2021:100).

One of the thesis' contributions is demonstrating negative AI impacts. As outlined in one of the feedback responses, there is a wide range of AI uses, resulting in a multiplicity of impacts that differ in their strength levels (Bublitz and Douglas 2021:5). This is why the definition of negative AI impacts is broad – negative AI impacts are changes negatively influencing individual and societal interests, as inspired by Smuha's definition of algorithmic harms (Smuha 2021a:4). Following Smuha's (2021a:4–5) definitions of algorithmic harms' spheres by distinguishing whose interests are affected, the societal impacts are interpreted as affecting social and collective interests, and individual impacts – are individuals' interests. Given the current status of legal protections focusing on individuals, the distinction between societal and group interests is not made. In this thesis, negative AI impacts can be summarised to include human rights violations, adverse AI effects on communities of children and workers, and specific AI applications resulting in negative impacts. Discussed AI applications are biometric recognition impacts and manipulative AI technologies. It has to be noted that discussions of algorithmic harms do fall under the negative impacts. The effects are termed as negative impacts instead of harms due to the diverging interpretations of whether harms are incurred across feedback responses and the challenges posed by societal impacts in proving that harm was inflicted.

1.3. Scope

Since the discussion of AI impacts across feedback responses takes place in response to the AIA, there is a fine balance to be maintained between discussing the impacts and conducting a legal analysis of the AIA. Thus, by default, there is a close link between the discussion of impacts and the specificities of the proposed regulation. That is why the scope of the close reading analysis is limited to revealing how respondents define the negative impacts and which particular use cases they mention to outline the effects that are not addressed by the Proposal. This specific focus comes at the expense of not fully engaging with the legal arguments and the specificities of formulation of the regulation's articles. This scope of engagement does not limit the quality of analysis centred around the negative algorithmic impacts overlooked by the AIA. Instead, it allows space for capturing the diversity of impacts without restricting the focus to several particular negative effects and then discussing the regulation's provisions in detail. An

example can illustrate the extent to which legal aspects are engaged. When criticism is raised about a particular article of the AIA, the positions or quotes of respondents on AI impacts are presented instead of providing legal analysis. Nevertheless, it has to be acknowledged that these negative impacts are expressed in response to the AIA.

Solutions proposed by the respondents are considered in close reading because of their direct relationship to impacts. Similarly to the extent of legal analysis of the AIA, these measures are discussed insofar as they help to understand negative AI impacts, given the limited scope of this paper. This decision was taken because the main focus of this thesis was limited to the analysis of impacts. Despite the limits of representativity across public calls for feedback and consultations (Niklas and Dencik 2021; Quittkat 2011; Vetulani-Cęgiel and Meyer 2021), the range of stakeholders of the call for feedback allows for capturing a relatively broad scope of impacts, given that there is a lack of cross-sectional approaches to algorithmic impacts (Parson, Fyshe, and Lizotte 2019:3–4). The solutions to the effects outlined are manifold and vary in their extent of challenging the EC's current approach. The assessment of overarching negative impacts and potential harms already offers a contribution as it establishes the major negative impacts and identifies the areas of agreement between different stakeholders.

1.4. Results

Unmitigated individual and social impacts by the AIA are identified across the five topics investigated: impacts of biometric technologies' use, manipulation and behavioural harms, effects on work conditions and workers' rights, impacts on children's use of technologies, and overarching human rights infringements across the four issue areas as mentioned above.

Discrimination, behavioural chilling effects, and surveillance are identified as negative biometric impacts. Manipulation ranging from captivation of attention, altering of behaviours and users' emotional states as well as its social effects, such as the impacts on political outcomes, polarisation, and misinformation, are noted as negative AI effects. Negative effects on autonomy, risks associated with algorithmic management systems and the absence of opt-out regimes are noted as individual impacts on working conditions and education. Monopolisation of education, labour market polarisation, subsequent inequalities and surveillance are identified as negative societal effects. AI influences on children are not as generalisable because they represent the viewpoint expressed in only one feedback response. Nevertheless, these negative impacts discussed in the topic of children outlined could be classified as associated with the effects discussed in biometric and manipulation topics. The infringements of human rights are noted across the topics discussed. The tackling of outstanding negative individual and societal impacts is then dependent upon the levels of determination by the European Commission, the European Parliament (EP), and the Council of the EU to challenge existing uses of AI.

H2 is partially validated with an overarching agreement across themes investigated using close reading on the need to put more resources into measuring AI impacts. Given that the discussions of negative AI impacts identified in H1 differ in their levels of certainty, the results from H2 are preliminary and indicate different levels of uncertainty. The uncertainty can be explained by the inconspicuous modes of functioning of certain technologies, such as in manipulation where technologies' functioning, subliminal techniques, is based on overriding rational control mechanisms of users. There is also a level of variation across different stakeholders and a level

of detail in discussions, resulting in different certainty levels of algorithmic impacts. The focus on discussions of technological uncertainty also reveals the lack of evidence supporting claims of technologies' capabilities. The scientific base of biometric categorisation and emotional recognition systems is questioned. Thus, the uncertainty is not limited to AI impacts; the asserted assumptions about biometric systems' efficiency are also questionable. The knowledge about algorithmic impacts can thus be improved by measuring AI capabilities.

2. Interdisciplinary State of Knowledge

Following the order of hypotheses, firstly, the state of knowledge of algorithmic impact studies is overviewed. In the first subsection, the primary guiding principle is developing an inquiry into cross-sectoral approaches to algorithmic impacts. In the following subsection, literature on the qualities of AI impacts and distinctions between individual and social levels are discussed. There, the cross-sectoral nature of approaches is not maintained. To inform the second hypothesis, the overview of challenges to measuring impacts is discussed. The process of uncovering AI impacts is contextualised in its informational sphere by outlining the asymmetries of informational power. Lastly, the use of topic modelling in policy fields. The developments in the topic modelling field and mixed method approaches from other disciplines are overviewed, outlining areas for improvement for policy studies' use of topic modelling.

2.1. Individual and Societal AI Impacts

Corresponding to the cross-sectional nature of the dataset investigated in this thesis, outlining major negative impacts unaddressed by the AIA, this state of knowledge subsection firstly overviews studies taking a cross-sectorial look at impacts, creating taxonomies and outlining distinctions between different impacts. Secondly, the literature on the qualities of AI impacts is presented by investigating the literature on AI impacts and algorithmic harms, which informs the analysis of algorithmic impacts induced from the analysis of feedback responses. This subsection provides examples of more granular AI impacts, thus expanding the mapping of negative AI effects.

It is beyond the scope of this thesis to overview the state of knowledge across algorithmic impacts fields and sectors; nevertheless, this subsection begins with a short introduction to the different types of approaches to the study of impacts. The presentation of this variety alludes to the explanation of the lack of cross-sectoral approaches to impacts. This division is noted as a challenging characteristic in the development of governance solutions (Parson, Fyshe, et al. 2019:2–4). This area is tackled by the cross-sectoral approach allowed by the data of feedback responses.

2.1.1. Diversity of Approaches to the Study on AI Impacts

There is a wide range of studies on algorithmic impacts. One of the branches of algorithmic impacts' study includes sectoral focuses on instances of algorithmic impacts, such as health (Whittaker et al. 2019), economy (Slaughter, Kopec, and Batal 2021), and the public sector (Bunnell 2021), amongst many others. Research on challenges in implementing any of AI principles (for an overview of AI principles, see (Floridi and Cowls 2019; Jobin, Ienca, and Vayena 2019; Schiff et al. 2020)) is another way in which AI impacts are explored and aimed to be mitigated (Arrieta et al. 2020; Mittelstadt et al. 2016; Obermeyer et al. 2019; Raji et al.

2020). The study of impacts can be carried by any theoretical framework, not limited to legal studies, lenses of feminism or critical theory (Parson, Fyshe, et al. 2019:4).

Another way AI impacts discussions occur in the public sphere is upon public attention and imaginary grabbing revelations and leaks about the impacts of widely used applications and lived experiences from the engagement with algorithms (Ganesh and Moss 2022). The Cambridge Analytica scandal (Susser, Roessler, and Nissenbaum 2019) and the recent Facebook Papers (Wells, Horwitz, and Seetharaman 2021) represent some of the well-known corporate scandals. From the public sector algorithmic impacts, the grade predicting algorithm based on historical school data in England developed in response to the COVID pandemic led high-school students to the streets to protest the Ofqual's grade attribution algorithm (Kolkman 2020). The Dutch Tax and Customs Administration had a range of scandals regarding its algorithms for identifying fraud (Heikkilä 2022; Vervloesem 2020), one of the algorithms was fined a 2.75 € million fine (Persoonsgegevens 2021). The already mentioned AIAAIC Repository can be referred to for a comprehensive list of impacts, which are categorised according to sectors, jurisdictions and technologies used, outlining the vast range of AI incidents.

As a result of this vast range of approaches to inspecting the impacts, this state of knowledge subsection below is focused on studies taking a cross-sectorial look at impacts, mirroring the cross-sectoral nature of the investigated dataset.

2.1.2. Mapping of Impacts

A multiplicity of taxonomies and categorisations of negative AI impacts outline different interpretations of individual and social impacts and provide a range of examples of negative AI effects. In this section, general approaches map the range of negative impacts in the fields of human rights, ethics, policy, and AI research.

The mappings of ethical and legal issues provide taxonomies of different granularities on impacts. Gasser et al. (2020) provide a taxonomy of legal and ethical challenges stemming from analysing digital tools used to mitigate COVID-19. The typology does not distinguish between levels of impact yet includes broader issues, such as public benefit, inequalities, and non-discrimination (Gasser et al. 2020:427). Rodrigues (2020) presents mappings of potentially infringed human rights principles according to AI legal issues and includes examples of vulnerable groups that can be affected. A comprehensive mapping of affected human rights does not distinguish but refers to individual, group and communal (children, women, migrants, older individuals, disabled people) and societal (free elections, social rights, public services) human rights principles that can be undermined by AI (Gasser et al. 2020:428). The noted mappings offer a broad range of ethical principles, human rights and issue areas of societal interests affected by AI.

Analyses of policy documents can be used to outline the major AI challenges signalled by policymakers, the direction that was intended by the approach taken by this thesis. Ulnicane et al. (2021) analyse 49 AI documents on AI policy in the EU and the US published between 2016 and 2018, focusing on how AI and associated concerns and benefits are framed. Ulnicane et al. (2021) capture concerns relating to the lack of AI definitions, cross-sectoral impacts in education and work, effects on human rights, values, norms and impacted AI ethics areas of privacy and safety. Power, geopolitical and political questions are identified as posing large-

scale challenges (Ulnicane, Eke, et al. 2021). Vesnic-Alujevic, Nascimento, and Pólvora (2020) analyse 21 EU policy reports, which were released between 2015 and 2018. By using categories of individual and societal realms set by Stahl, Timmermans and Flick (2017), Vesnic-Alujevic et al. (2020) conduct thematic analysis embedded in the grounded theory; thus, the authors form additional categories emerging in the EU policy documents to the categories developed by Stahl et al. (2017). Vesnic-Alujevic et al. (2020) classify nine issues into individual and societal categories. Fairness, diversity and good life, equity, accountability, transparency, responsibility, datafication, surveillance, and governance of AI are categorised as societal issues (Vesnic-Alujevic et al. 2020:3–4). In the meantime, privacy, data protection, autonomy and dignity are treated as individual issues (Vesnic-Alujevic et al. 2020:3–4). The investigated policy documents identify discrimination, inequalities, and chilling effects on behaviour stemming from surveillance and microtargeting (Vesnic-Alujevic et al. 2020). The distinction is made between microtargeting discussed under surveillance and challenges to political pluralism brought by personalisation of content and microtargeting, categorised as an AI governance issue (Vesnic-Alujevic et al. 2020). This distinction across uses (surveillance) and impact areas (governance) can outline the challenges in discussing the AI effects' issue areas since both categories refer to microtargeting. The two studies offer different granularities of analysis, with Vesnic-Alujevic et al. (2020) presenting an example of a more specialised categorisation exercise. This thesis is similar in its approach to mapping unmitigated AI impacts.

The societal impacts of AI are also a growing concern in the technical AI field. The analyses of the inaugural mandatory section on Broader Impact Statements of one of the leading AI conferences, Conference and Workshop on Neural Information Processing Systems (NeurIPS), provide a mapping of the main issues related to AI as perceived by the technical field. In analysing the most frequent words associated with societal impacts, privacy, biases, fairness, and safety topics are among the most discussed issues (Ashurst et al. 2021). In a qualitative analysis of the statements' sample, Nanayakkara, Hullman and Diakopoulos (2021:3–6) identified the following negative AI impacts: privacy and risks of increased surveillance; labour and displacement of jobs; environment and computational costs of their models; media and research's potential use for generation of fake news; bias and discrimination based on biased data. Despite the potential negative effects of impact statements (Prunkl et al. 2021:106–7), these impact assessments constitute an important exercise of reflexivity and provide identification of impacts as perceived by individuals developing AI theory and systems.

The taxonomies as mentioned above and mappings of impacts provide a range of negative AI impacts, ranging from human rights infringements to ethical challenges and the recognition of negative impacts across policy documents. The noted impacts guide the selection of topics for the guided close reading on negative AI impacts. This research contributes to the field by providing a case study of mapping major negative effects as revealed in the AIA feedback respondents' responses outlining issue areas unaddressed by the AIA.

2.1.3. Qualities of Algorithmic Impacts and Harms

This subsection covers literature on algorithmic harms and their qualities that outline the significance of tackling societal AI impacts.

Algorithmic harms are defined through their applications, qualities and areas of action. Tufekci outlines privacy, information asymmetries, algorithms' hidden influences and the lack of transparency as algorithmic harms (2015). Tufekci defines manipulation through its qualities: "algorithmic manipulation is neither public, nor visible, nor easily discernible" (2015:216). Marjanovic et al. (2021) propose reading of algorithmic harms as "algorithmic pollution" (2021), which is also distinguished by its qualities of being not visible and unrecognisable. Smuha (2021a:5) outlines the harms through their level of impact, outlining that "[societal harm] concerns harm to an interest held by society at large, going over and above the sum of individual interests" (Smuha 2021a:5). Differing conceptualisations of algorithmic harms allow for discussion of their qualities.

The opaque nature of algorithmic impacts is one of the qualities associated with algorithmic harms. This lack of visibility is associated with the potential for impacts amplification (Malik et al. 2022; Slaughter et al. 2021). Slaughter et al. (2021:1) maintain that AI systems "can simultaneously obscure problems and amplify them—all while giving the false impression that these problems do not or could not possibly exist." By presenting a case study of the Cambridge Analytica scandal and recommendation systems, Malik et al. (2022:189) note that "the technologies may distort the visibility and perception of social harms." Smuha (2021a) interprets this lack of visibility as a knowledge gap problem, where harm might not be noticed by those harmed. It is stated that "it may be even more difficult to demonstrate it and establish a causal link" (Smuha 2021a:9). In the meantime, Tufekci (2015) outlines the agile nature of algorithms and their constant development as resulting in difficulty noting the changes since there is no baseline for comparison. Different aspects of the lack of visibility outlined are complimentary. Constant tuning and optimisation of algorithms make identifying impact challenging, resulting in a challenging task to substantiate the harms incurred.

The accumulative nature of harms is another quality that adds to the reduced visibility of negative impacts. Li (2021:787) outlines the compounded aggregation of privacy harms for marginalised communities exposed to intersectional harms posed to privacy by increased digitisation due to the COVID-19. Harms stemming from the data reuse are interpreted as cumulative by Marjanovic et al. (2021). Smuha (2021a) identifies the constant rate of interactions with AI systems as another way accumulation takes place. The harms are caused by "often the widespread, repetitive or accumulative character of the practice" (Smuha 2021a:10). As a result, harms are accumulated through repeated engagements or accumulation through different sites of harm.

The harms discussed are societal in nature. Tufekci (2015) identified the level more specifically related to civic and political spheres. By comparing social media posts on the Ferguson protests of 2014 on Facebook and Twitter, Tufekci (2015) demonstrates the suppression of information on Facebook, thus presenting a case of harms in the civic sphere. Smuha (2021a) presents three examples of societal harm through breaches of equality, the rule of law and democracy. As revealed by the Cambridge Analytica leaks, political microtargeting and manipulation affect democracy (Smuha 2021a:6). By discussing biometric recognition, which results in both individual and collective harms, its impacts on equality challenge societal interest; thus, it is

also considered societal harm (Smuha 2021a). Smuha (2021a:10) notes that "societal harm will often manifest itself at a subsequent stage only, namely over the longer term." By outlining the social harms framework, Malik et al. (2022) analyse three case studies of Cambridge Analytica, the 2010 flash crash, and the Michigan Integrated Data Automated System (MiDAS) used in the sphere of public service, analysis of which resulted in the delineation of social harms qualities mentioned above.

What is apparent from these different approaches to harms is that the qualities of algorithmic harms are still in the developmental stage providing complementary explanations for the lack of visibility that makes harms hardly noticeable. The overview of algorithmic harms literature thus provides qualities of harms that make the identification of harms and proving their causal links extremely difficult. These difficulties outline the significance of addressing social algorithmic harms. This thesis contributes to the field by examining whether the AIA feedback respondents mention qualities of algorithmic harms in their discussions of negative AI impacts unaddressed by the AIA.

2.2. Discovery of AI Impacts and Informational Asymmetries

The study of impacts is directly related to the impacts' discovery process. This subsection discusses uncertainty associated with innovation, challenges at capturing AI impacts via governance tools and the context of power. The discussion outlines the main contributors to uncertainty, which inform the search of signals relating to difficulties in capturing and subsequently mitigating negative AI impacts, as considered under the H2.

2.2.1 Uncertainties and Discovery of AI Impacts

The innovation process is inherently linked to uncertainty (Stilgoe et al. 2013), which is even more prominent due to the revolutionary novelty, the pace of AI technologies development (Ulnicane, Eke, et al. 2021), and undetermined definitions (Nordström 2021). As Jalonen (2012:2) states, "uncertainty is inherent in innovation process." There are different components contributing to innovation uncertainties. Lane and Maxfield (2005) identify truth, semantic and ontological uncertainties. The assessments of the correctness of propositions correspond to truth uncertainty, while the semantic deals with the interpretation of the meaning of propositions, and ontological occurs when propositions cannot be created due to the lack of information (Lane and Maxfield 2005:9–10). Jalonen (2012) maps different types of uncertainties, such as technological, including the lack of details surrounding new technologies, social, political, and consequence uncertainties, among others. Parson et al. (2019b) indicate the social nature of processes through which technologies are created, resulting in social impacts. By noting the different aspects constituting uncertainties, uncertainties can be reduced by seeking more certainty in identified uncertainty areas.

From the perspective of uncertainties pertaining to AI, the knowledge of technologies' impacts and capabilities are identified as areas for improvement (Whittlestone and Clark 2021). There is a range of governance approaches directed toward increased responsibility as a response to uncertainties (Stilgoe et al. 2013). Standardisation is one of those governance approaches still in the making (Cihon 2019; European Commission 2020, 2021d), according to which the detailed requirements for AI systems in the AIA will be materialised. Impact assessments are another governance tool receiving increasing attention (Ada Lovelace Institute 2022; Ada

Lovelace Institute and DataKind UK 2020; Moss et al. 2021; Reisman et al. 2018), which are suggested as means to reverse informational asymmetries (Raji et al. 2020; Smuha 2021a).

The challenges of implementing impact assessments are plenty, discussion of which can reveal the challenges specific to uncovering AI uncertainties. Metcalf et al. (2021) maintain that the efficiency of impact assessments is reduced if the participation of diverse stakeholders, such as regulatory agencies, companies and academia, is not ensured. In this context, the lack of communal involvement in the impact assessments should be noted, given that the AIA does not accommodate for social evaluations of companies' self-assessments according to the requirements set for high-risk systems.

The issues with aligning impacts as close to harms pose a more fundamental challenge. Metcalf et al. (2021:735) provide an account for the challenges of impact assessments staying as close to harms as possible, "the impacts at the center of AIAs [algorithmic impact assessments] are constructs that act as proxies for the often conceptually distinct sociomaterial harms algorithmic systems may produce." The difference between harm and measurement of impact is explained by computational methods being less appropriate for estimating risks experienced by individuals and their groups (Metcalf et al. 2021:740). The possibility for phenomena of interest to not yet be captured by metrics is also noted in Thomas and Uminsky's (2020) argument for collecting qualitative accounts to capture the broad scope of potential user experiences. YouTube's algorithm is provided as an example demonstrating that short-term optimisation can lead to longer-term impacts not being measured, such as the impact on users' trust due to conspiracy theories being promoted by the YouTube algorithm (Thomas and Uminsky 2020).

The assessments ought to be adapted to their contexts and assessed for unintended consequences. Selbst et al. (2019) outline the importance of adapting impact assessments to their political and social contexts. Raji et al. (2020) outline the potential for ethical challenges raised by auditing biometric recognition systems. Companies' attempts to have more representative data potentially diminish underrepresented groups' privacy. The example of Zimbabwe's government agreeing to give access to its CCTV cameras to a start-up so it can diversify its dataset is presented (Hawkins 2018; Raji et al. 2020:4).

Impact assessments' efficiency in discovering and mitigating negative AI impacts is limited by the lack of inclusion of stakeholders, challenges of aligning assessments to capturing impacts and shortfalls in adapting the assessments to their contexts. This section outlines the uncertainties surrounding innovation and acknowledges difficulties in measuring and capturing AI impacts. The discussion of H2 contributes to the state of knowledge by providing a use case inspecting whether the components contributing to innovation and AI-specific uncertainties are acknowledged when discussing the main negative impacts outlined in response to the AIA.

2.2.2. Informational Asymmetries

As technology functions in and affects its social, political (Winner 1980) and cultural (Mohamed, Png, and Isaac 2020) contexts, the study of technologies is also embedded in them, which is why power and informational asymmetries are indicated.

The problem of private companies' asymmetrical power in shaping the discourses around societal algorithmic impacts can be interpreted as affecting the knowledge base surrounding impacts. Nemitz outlines accumulated power by the Big Tech companies, which is used to exert their influence through lobbying, owning means to where public discourse is taking place while

dominating the field of AI development and personal data collection needed to advance AI further (Nemitz 2018). The anti-competitive behaviours intended to maximise digital companies' profits and power exerted from the services provided by algorithmic means can result in abuses of that power (Pasquale 2015). Others identify market arrangements across the tech sector as oligopolies, which can explain the reason for societal impacts not being addressed (Ulnicane, Knight, et al. 2021), while others analyse these monopolies as "data-driven intellectual monopolies" (Rikap and Lundvall 2020:2) due to their use of public funding to conduct research while keeping the data private. This clash of corporate and public interest is noted beyond the short-term approaches and is already identified as a problem explored in the medium and long-term AI risks research (Baum 2017, 2020).

From the perspective of knowledge production surrounding AI and its effects, Abdalla and Abdalla (2021) compare Big Tech's funding to Big Tobacco in terms of their influence on research institutions, researchers, and Big Tech representation at conferences. By analysing the funding of faculty of 4 elite US universities, they demonstrate the conflict of interest in publications discussing the impacts of AI technologies. Even if their dataset is limited and the extent to which this trend can be extended can be challenged, the documentation of the Big Tech's relationship and funding of academia can be used to outline the significance of independent research. This context of power is significant in search of explanations for the reasons that make the untangling of algorithmic impacts challenging.

Nevertheless, it has to be noted that there are some steps taken to increase corporate transparency surrounding algorithmic impacts, as shown by Twitter publishing the results of an audit documenting higher content amplification by political right accounts than the left (Huszár et al. 2021), confirming the findings of Schradie (2019). Such transparency acts could be further analysed through the lens of performative transparency (Albu and Flyverbom 2019), situated in its political and social contexts (Felzmann et al. 2019).

The broader context of power and informational asymmetries favouring the Big Tech situates our relatively limited knowledge about algorithmic impacts and further signifies the importance of independent research on algorithmic effects. The informational asymmetries can contextualise the state of the evidence base used in the feedback responses.

2.3. Topic Modelling

Due to an increase in data generated and the accessibility of tools to extract and capture sets of existing data, the use of computational methods is on the rise with a wide range of applications across scientific fields (Baumer et al. 2017; Isoaho, Gritsenko, and Mäkelä 2021:301). Isoaho et al. (2021:301) noted that topic modelling (TM) is the most popular computational textual analysis technique, as observed across policy analysis journals. Given the popular use of TM in research, the following examples of TM use in policy studies are not comprehensive; yet, they present a wide range of different uses and interpretations of topic outputs across different policy fields. The uses of TM in policy analysis range from the analysis of the EP agenda themes and their development over time (Greene and Cross 2017), discourse analysis on EP trade policy (Jacobs and Tschötschel 2019) to the analysis of public perceptions on social distancing as expressed in surveys in the public health field (Ho et al. 2021). TM has also been used for modelling issue definitions around used nuclear fuel policy in the US Congress (Nowlin 2016), narrative analysis in energy policy (Debnath et al. 2020), and analysis of policy preferences and

their changes in communication due to changes in transparency in monetary policy (Hansen, McMahon, and Prat 2018), amongst other uses in the economics field (Gentzkow, Kelly, and Taddy 2019). In the technology policy field, TM has been used to analyse discussions across public consultation responses to the EC consultation on roaming regulation before publishing the respective draft proposal (Alves et al. 2021). When it comes to AI policy, TM has been used to discover topics across national AI strategies (Papadopoulos and Charalabidis 2020) and legal journal articles (Rosca et al. 2020). This overview demonstrates that TM is widely adopted in policy research across a wide range of uses, such as thematic analysis for classification and identification of policy issues to discourse analysis and problem definition.

Despite the adoption of TM approaches in different branches of policy, there is a lack of established conventions for TM analysis in policy research (Isoaho et al. 2021). The use of TM in the policy field can be criticised for lacking the rigorous engagement with model evaluation, parameter setting and interpretations of the TM outputs (Isoaho et al. 2021:306), which is why output's validity can be challenged (Grimmer and Stewart 2013).

Mixed method approaches can be used to balance out the limitations of TM techniques that treat the words outside of their contexts due to the bags of words assumption (Eickhoff and Wieneke 2018; Isoaho et al. 2021). Baumer et al. (2017), in their comparison of TM and grounded theory, noted the potential of combining the approaches, despite epistemic differences between positivist and interpretivist traditions (2017:1406). Both methods share a common target area of thematic patterns research, both methods are embedded in data, and lastly, both approaches are iterative (Baumer et al. 2017). By combining TM with grounded theory, TM's outputs can be improved by going beyond computational metrics of performance; thus, bringing TM closer to improving the model performance in terms of the interpretation of the outputs (Baumer et al. 2017:1407). Nelson (2020) further develops the combination of the two approaches as computational grounded theory. The ways computational methods can be used to select the most representative documents of the topic are outlined (Nelson 2020). By combining TM with qualitative methods, the research readers "can trust that when a quote is chosen as an example of something, it is not an outlier but is indeed representative of some theme in the text" (Nelson 2020:26). Topic distributions for the texts inform "[c]omputationally guided deep reading" (Nelson 2020:26), providing a researcher with guidance on which topics are present in the inspected text.

As a result, the method of the thesis benefits from mixed method approaches combining TM with qualitative methods to expand the validity and better situate the topics in the contexts from which they were induced by TM. In this way, this overview does not identify the gap in knowledge on TM applications. Nevertheless, the mixed methods approaches are not such a prevalent practice in the policy field (Isoaho et al. 2021). As a result, the contribution of the chosen method in the policy does not attempt to be groundbreaking; instead, it aims at the strategic adoption of best practices of TM mixed method approaches used by other fields.

3. Design, Methodology, Data and Validation

3.1. Mixed Methods Design

As discussed in the state of knowledge section, a mixed method of combining TM with close reading allows one to consider the diversity of themes across the corpora without limiting the initial analysis by the topics of interest to a researcher. The mixed methods approach assumes that the qualitative and quantitative methods are compatible (Eickhoff and Wieneke 2018). As outlined by Baumer et al. (2017), TM and qualitative analysis based on grounded theory can be considered to be complementary. As a result, TM is used to discover issue themes across the feedback responses to the AIA inductively, which are identified by labelling the most frequent words from the topics generated. The subset of topics is selected for close reading, which is guided, as suggested by Nelson (2020), by topic weights probabilities to identify the most significant feedback responses for the themes chosen. Guided close reading allows one to tackle the question of what the main negative AI impacts are unmitigated by regulation.

Computational TM techniques discovering topics are chosen as an initial step for analysing feedback responses due to the method's utility in uncovering latent topics across the corpora without presupposing the themes for analysis. Compared to non-computational methods of thematic analysis, such as qualitative content or thematic analysis, the costs of analysis are reduced due to TM's computational base (Grimmer and Stewart 2013). TM allows the analysis of a large dataset without requiring sampling or subsetting the data, as it is usually done in qualitative studies. TM allows the analysis of the feedback responses as they are, with only a limited extent of modifications. TM methods which discover latent themes allow getting rid of biases of analysing the text through the researchers' lens, which can impact the discovery of themes (Nelson 2020).

Inspired by computational grounded theory (Nelson 2020), the relationship between TM and close reading for this thesis is dynamic. According to the guided close reading, the interpretation of topics and their themes is adapted. With the guided close reading, the diversity and plurality of perspectives on more specified algorithmic impacts are discovered within the theme, improving TM models' interpretation. It also allows one to validate the generated model. As a result, guided close reading with a focus on contextualising themes discovered by topic modelling is used as part of the sequential design in the analysis of a selected subset of themes revolving around negative AI impacts.

The choices made when developing a TM model evidently influence the associated objectivity surrounding the discovery of themes (Baumer et al. 2017); thus, the choices are presented in the following methodology section.

3.2. Methodology

In this section, the method for negative AI impacts analysis as expressed in the feedback responses to the AIA is presented. The mixed method approach is used to find topics generated by the quantitative data analysis technique TM. In this section overview of the theory behind TM, reasons for choosing TM and modelling choices in making the TM model are presented. The approach to the close reading on negative AI impacts issues is detailed. The limitations of the method are outlined, and data is presented.

3.2.1. Topic Modelling

TM is a statistical model used to analyse text data that constitutes a dataset, also known as corpus, in order to find its semantic structure, which is constructed from different topics (Blei and Lafferty n.d.). There are multiple topic modelling techniques, according to which their particular ways of estimating the topics would change the explanation of how topic modelling functions. According to the Latent Dirichlet allocation (LDA) modelling, the documents are considered to be constructed from a mixture of different latent topics (Blei, Ng, and Jordan 2003). These topics are represented by different words distributions. In other words, documents are made from probability distributions of different combinations of topics, constructed from probability distributions of lists of words (Jelodar et al. 2019). The repeated iterations of combining texts based on probable topics are run to maximise topic and document probabilities (Jelodar et al. 2019). By reading the list of terms from topics generated, a theme of the topic can be usually inferred (Jelodar et al. 2019). LDA models produce topics with distributions of lists of words associated with the topic without labelling the topics. Meaning that the objectivity of the model is still constrained by the researcher's interpretation of the topics and the method chosen to validate them (Nowacki, Monk, and Decoster 2021), which is why emphasis is placed in this section to explain the choices made when selecting and creating the model.

It is important to acknowledge this multiplicity and that the choice of a model will depend on the question the model is tackling, which guides the choice of modelling technique (Grimmer and Stewart 2013; Isoaho et al. 2021). The LDA model is chosen for this thesis. LDA can be chosen for analysis when documents contain different numbers of topics that vary across documents or when most documents share a portion of topics (Isoaho et al. 2021). As noted by Isoaho et al. (2021), LDA modelling can be less effective when applied to EU policy documents, where a few topics are represented across corpora and there are a few niche topics. The most frequent topic-specific stop words were removed to address the challenge of disproportionate topic distribution of several topics across most feedback responses. This approach to the most frequent words is still endorsed by those questioning the utility of stop word removal (Schofield, Magnusson, and Mimno 2017). By removing overarching words, such as "AI," "regul," "propos," the distribution issue is addressed, and LDA thus can be applied to the AI Act feedback analysis.

The choice of LDA was also considered with respect to other models. The author is familiar with non-negative matrix factorisation (NMF) models that have been shown to provide more coherent results than LDA in modelling corpus when there are more niche topics (O'callaghan et al. 2015) and in modelling topics discussed in the European Parliament (Greene and Cross 2017), which is why topic distribution issues were discussed and addressed previously. The choice to select LDA models is based on the length of the documents since LDA models are used for longer texts (Guo, Lu, and Wei 2021), while NMF and structural topic modelling (STM) are better suited for shorter texts (Yan et al. 2013). STM is a model that treats metadata about texts as covariates (Egami et al. 2018). In the metadata available for the AIA feedback responses, feedback responses' categorisation according to sectors, the size of an entity and the country of origin could be used to model the topics. However, the length of feedback responses ranging from a page to 82 pages prevents one from choosing STM models, which function better on shorter documents.

3.2.2. Guided Close Reading

The topics relating to AI impacts are chosen for guided close reading. The criteria for selecting close reading topics were set aiming to capture the topics discussing negative AI impacts most accurately. After the initial reading of the 20 most probable words across topics, two categories of words were formed as selection criteria, one signalling impacts and another identifying more specific impacts as identified by studies covering a range of AI impacts. The threshold of the number of words' recurrences is then selected. As with any threshold, it is a choice where each side of the line has its own set of advantages and disadvantages. The threshold was set in the context of inspecting the topics, most probable words and topics' labels, which were induced based on the 3 most representative documents of the topic. The threshold was set while balancing the time resources needed to conduct guided close reading and interpreting which topics capture the negative impacts. This selection process could be challenged as omitting neglected algorithmic impact areas. However, the neglected areas might not be captured by TM by default due to the set threshold of minimal frequency of words, which improves the model's accuracy and reduces the running time while not preventing one from discovering the major negative impacts unaddressed by regulation.

After the themes for close reading are selected, the matrix with distributions of documents across topics is used to guide the selection of documents for close analysis. The top 10 documents with the highest probabilities of the topic of interest are selected for analysis, inspired by Nelson's demonstration of computational grounded theory (2020:28). The probability can be understood as if the probability of a topic is above 0.5, which means that most of the text in the document is predicted to be on the theme of that topic. This approach to choosing the documents that have the highest distribution of the topic of interest allows to subset the documents analysed to those in which the topic plays a relatively significant role. What is meant by relatively is that document probability rankings across the topic are relative to that document in terms of other topics inside the document. The overall sum of topics' probabilities inside one document equals 1.

Given that topic weights are assigned relative to other topics inside the document, there might be documents discussing the topic of interest that are not analysed in close reading. The reasons for such documents not appearing in the top 10 list could be because they might be longer documents, and there might be a lower number of words devoted to the topic on a particular impact. Thus, the topic would be relatively less recurring and represented than other topics, which is how such a document, even if containing that particular topic, could have evaded being engaged with close reading.

During close reading, there were instances where additional themes of impacts were observed, such as environment and human relationships. The choice to not include them in the discussion was based on the interpretation of topics generated as signalling the most significant negative impacts and whether the discussion of those other impacts was related to the topic, such as biometric recognition technologies. Due to the modelling choices that will be discussed below, the words associated with the topic had to occur at least 5 times to be included in the model. This means that if the issue and words associated with it were not present in the topics produced, the topic did receive less attention.

3.3. Data Collection and Processing

3.3.1. Data Collection

The feedback responses (also referred to in this work as texts, documents, and responses) to the AIA were manually downloaded from the EC website hosting the responses (European Commission 2021a). Feedback responses included attached PDF format files were downloaded. When respondents did not upload the PDF feedback responses, the PDF print of the response page was extracted. 304 responses were submitted to the European Commission, of which 303 were valid. As a result, 303 PDF files constituted the dataset for analysis.

Cleaning of data submissions consisted of removing one empty submission and extracting submissions written in other languages than English. Since the dataset was collected manually, one empty feedback form containing only the word "test" was deleted. Regarding submissions written in other official EU languages, it was chosen against the translation of texts written in languages other than English. Despite the advances in neural machine translations, the translation would bring some level of uncertainty in the quality of translations. The author could not control the quality of translations, given that eight additional languages were used in responses. Thus, since the main aim was to analyse texts as written and intended by their authors, texts written in other official EU languages were excluded. In addition, another reason for removing documents written in other languages was not to reduce the accuracy and precision of the model. TM is based on analysing the recurrences of words and how they are related to other words across the documents, which is why adding other languages would result in less coherent topics. As a result, it is optimal for the final dataset to be restricted to the English corpus only.

The responses written in languages other than English were captured by looking at the list of the least frequent words. Those non-English words were inserted into the neural machine translator *deepL* and identified as belonging to those languages. After other than English languages were identified, the translations of the words "AI" and "Commission" were searched in the collected feedback responses. These non-English responses were deleted from the dataset. The process of looking for non-English language words was repeated until words in languages other than English were not present in the list of words in the corpora. The responses written in two languages (including English) were maintained in the final dataset by cropping the rest of the document that contained other official EU languages.

Text duplicates were present in the dataset; for instance, Digital Courage (2021) uploaded EDRi's (2021) submission, and two different labour unions (GEW 2021; ČMOS 2021) uploaded the same ETUCE's (2021) response. It was decided to keep these submissions because they are responses that differing stakeholders chose to upload, assuming they chose to upload the position as best representing their position. As a result, the copies of texts were not considered as duplicates because different stakeholders uploaded them.

After data entries cleaning by language, the dataset for analysis contains 266 feedback submissions, constituting 87% per cent of the valid responses submitted to the EC.

3.3.2. Corpus Pre-processing

Another crucial step in increasing topic model precision is cleaning the text corpus (Isoaho et al. 2021). The following order of corpus pre-processing steps was implemented as following recommendations set by (Grimmer and Stewart 2013:272–73):

1. All capital characters are transformed to lowercase to treat the words with different capitalisation levels as the same word.
2. URLs were removed. A small number of entries had footnotes and references. The choice was made not to delete the footnotes since they were used to understand the knowledge base used by the respondents to support their identification of the unaddressed AI impacts.
3. Common stop words were removed from the English stop words library.
4. Deleted punctuation since LDA is based on bag of words assumption, meaning that the order of words is not significant, which is why punctuation is removed.
5. Numbers were removed.
6. The words were stemmed to treat the words of the same stem as the same word, "increas", which can contain words, such as "increased", "increasing", "increase."
7. The most frequent AI policy feedback responses specific stop words were removed. Even when challenging the effects on validity by the removal of stop words (Schofield et al. 2017), the removal of the most frequent terms is suggested. The words with the lowest sparsity were identified up until being represented across 80 % of the documents. The words identified in terms of sparsity were then compared to the most frequent words list of 100 words. The resulting words were deleted.

3.4. Topic Modelling

The TM model was built using RStudio open-source software. After the pre-processing steps were concluded, the document-term matrix was created using the *TermDocumentMatrix* function from the *tm* package. The minimum frequency of words across the corpora was set to 5. The *LDA* function from the *topicmodels* package was applied to create the model. *Gibbs* sampling method was used, seed was set to 9123, and 500 iterations were run. Below is the discussion of the selection process for the main model parameters.

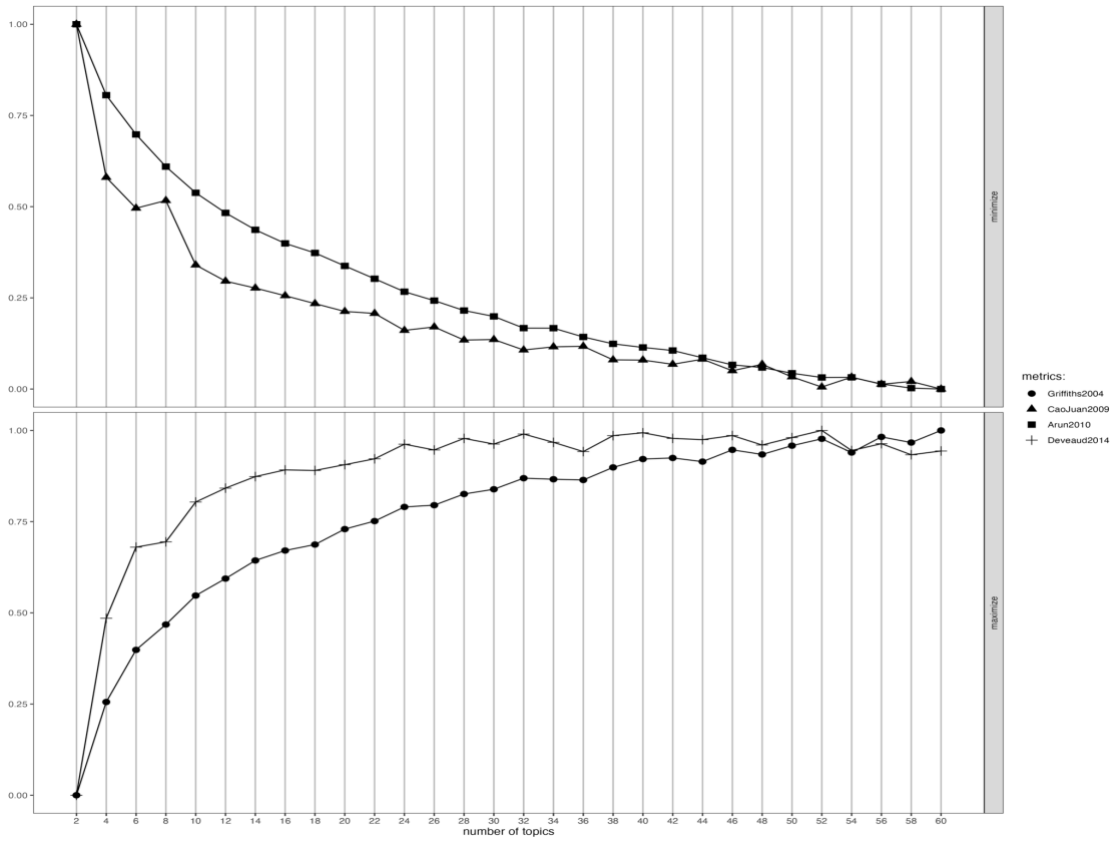
3.4.1. Selecting the Number of Topics

The discussion around the selection of the number of topics in the light of using TM for analysing problem definitions was conducted by Nowlin (2016), who maintained that the selection of the number of topics, *K*, is dependent on hypothesis and theory. Thus, in Nowlin's case of mapping issue definitions, *K* is assumed to be situated at the lower end to capture different approaches to the issue. In this thesis, the TM is used to capture different issue areas. As a result, the assumption is made that the *K* ought to be higher.

The number of topics *K* is firstly computed statistically using the *FindTopicsNumber* function from the *ldatuning* package. The range of topics investigated was set from 2 to 60, and the function was run at the intervals of 2 topics. The function's output is represented in Figure 1

below. The function provides the interpretation of the metrics as to which the selection of *K* ought to be maximised and minimised.

Figure 1: The search for the statistically optimal number of topics across four *ldatuning* metrics



Several numbers of topics were tested, maintaining the guiding principle that TM outputs should be granular enough to discover different themes of negative AI impacts. Models with *K* set to 16, 24, 34 and 52 were run. It can be seen in Figure 1 that the performance of the TM model is steadily increasing until 10 topics, after which the rate of improvement decreases. The diminishing rate of improvement is particularly noticeable after reaching the local maximum at 24 topics. The tested numbers of topics were compared according to their resulting outputs, whether the generated topics and their most probable words can be associated with impacts and risks, and whether there is not too great of an overlap between the topics, which is more likely when the number of topics is increased with reduced legibility for model interpreters. The trade-off is thus faced when selecting the number of topics that have higher granularity and coherence of topics. *K* is set for 24 because the local maximum is reached at 24 while the rate of improvement of metrics is still decreasing slower and because 24 topics represent granular enough outputs for identifying topics where negative impacts are discussed.

3.4.2. Model Validation

The absence of a standard TM validation procedure (Heidenreich et al. 2019) is explained by the need to specifically adapt the model validation to each model (Boussalis and Coan 2016). As a result, there is a multiplicity of means to validate the model. Some of the means are focused on output interpretation, parameter choice and the extent to which the model captures the area of investigation (Isoaho et al. 2021), which by others are categorised as predictive, semantic, and statistical approaches to validation (DiMaggio, Nag, and Blei 2013).

As the number of topics, *K* is one of the main parameters that can affect the quality of the model (Isoaho et al. 2021), the selection of the *K* was informed by statistical metrics to select the most statistically appropriate number. It was verified by testing models with different *K* and selecting the number by balancing the trade-off between topics' granularity and coherence.

For unsupervised models, such as LDA, the validation of the model focuses on its outputs (Grimmer and Stewart 2013:286). In this case, semantic validity is assessed by looking at the 20 words with the highest probabilities of being associated with the topic assessed. As Isoaho et al. (2021:306, 312) outline, the top words should be assessed according to how they are represented across documents. As a result, the top 3 documents according to the topic probabilities are read to validate the topic. Topic labels are changed if the initial label does not correspond to the theme in which the top words are used, following the approach outlined by Nelson (2020). The selection of documents with the highest topic probabilities guides the close reading; thus, the most representative responses are selected. This allows one to test the significance of the most probable words of the topics and analyse the themes in which they are applied.

4. Analysis

The following steps form the analysis. Firstly, the overarching 24 topics generated with topic modelling are presented, and the selection of the subset of topics for the analysis of the algorithmic impacts is discussed. Secondly, the first hypothesis is tested by discussing the negative AI impacts presented in the top 10 documents of selected topics analysed through close reading. Thirdly, the second hypothesis is examined by discussing the levels of certainty expressed across the investigated documents and their positions regarding the need for more research on impacts.

4.1. Topics across Feedback Responses to the AIA

The corpora are analysed by setting the number of topics, *K*, to 24. The results from the topic modelling are represented in Figure 2 on the next page. The topic words with the highest probabilities informed the assignment of the topic label. Following the method outlined in the previous section, three documents of each topic were consulted to verify if the topic labels were induced correctly and changed accordingly if needed.

TM topics that contain at least two words from the categories below across the 20 most probable words list were selected for guided close reading. One group refers to words signifying impacts, such as "harm", "social," and "impact," directly referring to AI impacts. The other is constituted from stems of words referring to specific AI impacts outlined in the literature review, such as "work," "manipulation," "politics," "surveillance," and "environment." The selection criteria for subsetting the topics for close reading are thus embedded by the target area of impacts represented by words signalling impacts and impact specific words identified by the state of knowledge review.

Topics 4, 6, 12, 13 and 16 were chosen for analysis. In the following section, results from the close reading of the top 10 documents of the selected topics outline negative AI impacts unaddressed by the AIA.

Figure 2. TM topics and their top 20 words stems in the order of decreasing probabilities

Topic ID	Topic label	Top 20 word stems in order of decreasing probabilities
1	Model design	model, algorithm, process, method, risk, explain, explan, perform, level, case, valid, intern, busi, evalu, predict, oper, technic, design, decis, document
2	Ethics operationalisation	ethic, process, learn, principl, research, make, practic, bias, tool, machin, creat, publish, train, address, chang, mitig, design, set, model, knowledg
3	Security of services	servic, process, custom, control, law, protect, purpos, chain, secur, gdpr, model, direct, cloud, review, potenti, offer, comput, activ, oper, stage
4	**Impacts of biometric technologies' use**	prohibit, right, person, risk, emot, recognit, public, fundament, **social**, access, natur, biometr, law, practic, state, **impact**, protect, oblig, purpos, group
5	Research investments	research, plan, coordin, europ, respons, invest, recommend, network, state, concern, innov, fund, creat, agre, global, area, address, challeng, excel, effect
6	**Manipulation and behavioural harms**	person, influenc, **manipul**, control, problem, behaviour, right, peopl, suggest, target, decis, **harm**, psycholog, case, signific, affect, mean, specif, direct, physic
7	Healthcare for patents	health, healthcar, patient, access, digit, care, potenti, profession, safeti, qualiti, innov, increas, solut, specif, public, framework, level, improv, ethic, sector
8	Autonomous vehicles	consum, vehicl, access, market, insur, econom, highrisk, independ, aia, legisl, area, **harm**, car, public, competit, manufactur, autom, servic, bundesverband, connect
9	Human rights protections	right, control, gdpr, fundament, **impact**, adm, risk, access, process, protect, human, law, person, legal, privat, individu, subject, interest, univers, app
10	Industry standards	standard, industri, standardis, specif, market, harmonis, europ, legisl, common, smes, conform, technic, adopt, request, open, brussel, certif, qualiti, support, issu
11	Procedural terms of submissions	opinion, organis, view, submit, refer, countri, issu, regist, union, transpar, legal, ethic, initi, type, user, infolaw, accuraci, effect, guarante, remov
12	**Work conditions and workers' rights**	educ, public, tool, risk, **work**, nation, employ, equal, sector, digit, **social**, worker, govern, teacher, trade, union, right, protect, skill, collect
13	**Children and impacts of their use of technologies**	children, digit, design, **social**, peopl, onlin, user, time, media, strategi, year, **environ**, young, age, child, human, right, creat, world, chapter
14	Safety of medical devices	medic, devic, manufactur, mdr, aia, risk, bodi, notifi, softwar, manag, market, addit, recommend, document, ivdr, product, perform, exist, legisl, test
15	Financial scores	financi, risk, credit, institut, servic, bank, suggest, score, creditworthi, refer, manag, approach, definit, techniqu, high, firm, process, supervis, consum, appli
16	**Harms and human rights**	aia, right, risk, fundament, **surveil**, human, **harm**, limit, biometr, peopl, enforc, law, **impact**, process, high, societi, potenti, mass, deploy, context
17	Human right protections	right, fundament, protect, law, legal, enforc, highrisk, individu, risk, public, trustworthi, subject, effect, list, prohibit, practic, **harm**, transpar, approach, scope
18	Innovation and sandboxes	compani, innov, sandbox, busi, user, member, cost, complianc, associ, product, risk, high, digit, make, state, gdpr, creat, competit, industri, small
19	Development of standards	support, innov, approach, standard, framework, govern, global, code, encourag, risk, trust, recommend, build, transpar, respons, solut, promot, exist, market, sourc
20	Biometric systems prohibition	biometr, identif, remot, public, human, individu, highrisk, effect, recommend, prohibit, enforc, except, content, affect, person, access, recognit, realtim, algorithm, bias
21	Operationalisation of high-risk requirements	highrisk, user, oblig, risk, case, clarifi, market, set, conform, deploy, specif, product, practic, provis, safeti, error, train, monitor, definit, person
22	Definitions and scope of high-risk requirements	definit, highrisk, approach, risk, safeti, product, scope, consid, list, legisl, softwar, relev, defin, purpos, refer, oblig, case, broad, exist, intend
23	Legal governance structures	legal, state, process, implement, nation, term, order, level, addit, compet, govern, board, member, technic, establish, avoid, text, set, union, area
24	Testing	test, decis, autom, transpar, predict, law, individu, person, organ, crimin, bias, decisionmak, human, justic, risk, standard, make, outcom, analysi, profil

4.2. Guided Close Reading of Selected Topics

In this analysis section, the negative AI impacts are outlined as they were discussed in the top 10 documents across selected topics. The top 10 highest-ranking documents are chosen according to the ranking of topic probabilities of the topics of interest across 288 documents analysed with topic modelling. Appendix 1 contains the list of data analysed for close reading, which serves as a bibliographic reference for referred documents. It includes 50 feedback responses, 10 documents per topic, associated responses' in-text abbreviations and their topic probabilities are provided.

Below are the discussions on biometric recognition, manipulation, and AI impacts on working conditions and children. Human rights topics are not discussed separately because the top 3 documents with the highest probabilities across the topic are from topic 4, and five of the top ten documents were analysed in the rest of the topics selected for the close reading. As a result, the remaining five human rights documents were attributed to the other highest topic probability ranking amongst the analysed 4 topics. As a result, in each subtopic, implicated human rights are discussed.

The analysis of chosen topics outlines the negative impacts discussed in the feedback responses. Each section is introduced with an overview of the respondents' sectors. Biometric recognition and manipulation sections include a presentation of AIA articles shaping the discussions of feedback responses; meanwhile, across work conditions and AI impacts on children's topics, impacts are discussed more generally as reflecting the analysed feedback responses. There is a slight variation in the organisation of the following sections dictated by the different characteristics of negative effects discussed.

4.2.1. Impacts of Biometric Technologies' Use (Topic 4)

Topic 4 contains discussions surrounding biometric identification technologies and uses of biometric data, such as biometric categorisation and emotion recognition systems. The technologies used by law enforcement and border control are also discussed. There is an agreement across the top 10 topics' documents, maintaining that biometric systems pose risks to human rights. NGOs submitted seven responses from the ten analysed with close reading, and the academia submitted the rest.

Definitions of biometric technologies' use. Before going into details about biometric impacts, it is essential to indicate what data belongs to biometric data and for which purposes biometric systems are used. According to Article 3 of the AIA, which provides definitions of the AIA regulation, biometric data, emotion recognition and biometric systems are defined as:

> (33)'biometric data' means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data;
>
> (34)'emotion recognition system' means an AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data;
>
> (35)'biometric categorisation system' means an AI system for the purpose of assigning natural persons to specific categories, such as sex, age, hair colour, eye colour, tattoos,

ethnic origin or sexual or political orientation, on the basis of their biometric data. (European Commission 2021c)

Most of the investigated responses challenge these definitions. For the purpose of this thesis, it has to be noted that according to the definition of biometric data, biometric data that does not allow for individual identification is not considered biometric data.

Law enforcement's use of real-time biometric identification systems is prohibited except for determined cases, Article 5.1.(d)(i). Transparency requirements for emotion recognition systems are laid down in Article 52.

Negative AI impacts of biometric technologies' use

All feedback responses investigated in close reading noted the negative impacts and infringements of human rights. Topic 4 encompasses negative impacts, such as chilling effects on behavioural changes and fundamental rights, surveillance risks, human rights, and discrimination.

Chilling effects on behaviour. Both biometric categorisation and emotional recognition systems were identified as potentially producing chilling effects on individual behaviours and human rights (ALLAI 2021; Amnesty International 2021:2; Hildebrandt 2021). Due to the process of being identified and analysed by biometric categorisation and emotional recognition systems, privacy, autonomy, and identity are affected, impacts which are summarised as intruding psychological integrity (ALLAI 2021:13). The chilling effect of being exposed to biometric systems can result in individuals altering their behaviours, "As a psychological 'chilling' effect, people might feel inclined to adapt their behaviour to a certain norm, *which shifts the balance of power between the state or private organisation using facial recognition and the individual* [emphasis added]" (ALLAI 2021:13). Access Now (2021:10) exemplifies this chilling effect on behaviour by referencing Canon's biometric systems example, where only smiling employees are allowed to access particular spaces inside their offices. They extend this example to a hypothetical one, where employees showing a determined level of satisfaction could receive bonuses while those scoring lower in the perceived emotional state could be disciplined (Access Now 2021:10). As a result, this chilling effect on changes in behaviour can take multiple forms, from modified behaviours to constraints from taking certain actions. These, in turn, can result in different strengths of impact, from the inability to access specific spaces to more significant effects, such as restricting opportunities for bonuses.

Chilling effects on human rights. The second realm of chilling effects affects fundamental rights. NGO Access Now noted these chilling effects on rights, "the chilling effect they create on freedom of expression and freedom of assembly and association" (2021:17). The chilling effect can have a "detrimental effect on people's ability to protest and to fully enjoy public space" (Access Now 2021:17). In another way, this chilling effect can result in "people no longer partaking in peaceful demonstrations" (ALLAI 2021:14). The Centre for Commercial Law, University of Aberdeen (CCLUA) demonstrates how the non-real-time use of biometric identification could affect human rights, despite the AIA provisions limiting the use of real-time biometric recognition systems. Human rights of already mentioned freedom of association and assembly, "and more in general the founding principles of democracy" (CCLUA 2021:8), could be affected, such as after a political protest. Thus, chilling effects on fundamental rights

are impacting individual rights, resulting in infringement of societal rights and their limited exercise, which has societal consequences.

Surveillance. Related to chilling effects, risks of surveillance receive considerable attention. Research institution the Center for AI and Digital Policy (CAIDP) (2021:3) recommends banning biometric systems used for mass surveillance. EDRi (2021:18), too, notes the danger of surveillance: "Biometric categorisation often forms the technical foundation of other forms of biometric data processing which can lead to mass surveillance." Not only surveillance by states and law enforcement is considered, Amnesty International notes that surveillance is also problematic when used by private entities: "We have seen a surge in biometric recognition applications in workplaces, in recruitment and human resources and in commercial settings, leading to corporate surveillance and the widespread use of techniques that constitute a threat to our human rights" (2021:3). As it can be seen, surveillance is an acknowledged potential risk unaddressed by the regulation, which nevertheless takes steps to limit the biometric systems used by law enforcement. Regardless, as outlined by Amnesty International, corporate surveillance poses risks too and can affect multiple spheres of individuals' lives, such as in workplaces and commercial spheres.

Discrimination. Investigated feedback responses note the risks of discrimination associated with the use of categorisation systems and subsequent infringements of the right of non-discrimination, such as Article 21 of the EU Charter of Fundamental Rights (CAIDP 2021:5) as reasons for banning biometric categorisation (Amnesty International 2021:1; Hildebrandt 2021:6). As noted by EDRi, "By definition, biometric categorisation is a process that seeks to put people into (often arbitrary, discretionary and stereotyped) boxes, and then to make predictions or decisions about them on that basis" (2021:18). By mentioning the historical uses of injustice and oppression based on categorisations of people, EDRi (2021:18) maintains that biometric categorisation uses are "exceptionally hard to justify," because of discrimination and threats to equality. Furthermore, Access Now (2021:10) notes with high certainty that biometric systems could have discriminatory impacts, "systems will lead to discriminatory impacts on already marginalised and racialised groups." Access Now (2021:11) refers to studies demonstrating that Black people are assigned more negative emotions (Rhue 2018) and that existing stereotypes are reproduced by algorithms (Rhue 2019; Access Now 2021:10). As a result of biased algorithms, negative impacts disproportionately affect individuals outside of the prevailing categories, thus, furthering discrimination (Amnesty International 2021:1-2; Hildebrandt 2021:6).

Human rights. Additionally to chilling effects on exercising human rights, the aforementioned algorithmic impacts of biometric systems can be summarised as violating privacy and its categories as enshrined by the ECHR Article 8 (ALLAI 2021:13), freedom of expression, equality, right to association (Amnesty International 2021:2), movement (CAIDP 2021:3), protest and non-discrimination (Access Now 2021:10). Other infringements are also documented, such as the right to an opinion, to express one's identity (CAIDP 2021:5), and freedom of thought (Access Now 2021:10). In this way, human dignity (Amnesty International 2021:2; EDRi 2021) and expectations of anonymity (CAIDP 2021:3; Hildebrandt 2021:4) are negatively affected.

4.2.2. Manipulation and Psychological Harms (Topic 6)

Topic 6 focuses on AI systems circumventing users' behavioural control mechanisms leading to manipulation. Documents discuss difficulties AI systems users face in proving that psychological harms were inflicted. A range of manipulative influences at both individual and societal levels is discussed, together with accompanying human rights violations. A diversity of stakeholders expressed their views on manipulation: four NGOs, three business organisations, one academic institution, and two other types of stakeholders submitted their responses.

Definition of psychological harm. Article 5 of the AI Act outlines prohibited AI uses. As it relates to AI systems affecting psychological states, Article 5.1(a) prohibits:

> [T]he placing on the market, putting into service or use of an AI system that deploys *subliminal techniques beyond a person's consciousness* [emphasis added] in order to materially distort a person's behaviour in a manner that *causes or is likely to cause that person or another person physical or psychological harm*. (European Commission 2021c)

Another prohibition relating to systems resulting in behavioural impacts prohibits the exploitation of vulnerabilities of vulnerable groups. Article 5.1(b) prohibits:

> [T]he placing on the market, putting into service or use of an AI system that exploits any of the *vulnerabilities of a specific group of persons due to their age, physical or mental disability* [emphasis added], in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm. (European Commission 2021c)

These two provisions inform the context in which the top 10 feedback responses having the greatest topic weight of topic 6 on manipulation are situated.

Negative AI impacts: manipulation

Even if the AIA prohibits AI uses that result or can result in psychological harms, the regulation does not address subsequent negative psychological impacts due to the functioning mechanisms of AI systems applying manipulation. These systems bypass behavioural control, which is one of the reasons they are so effective.

The opacity of manipulation. The changes brought by these systems are hard to pinpoint, making the process of proving that harm was inflicted almost impossible. Multiple responses emphasise these difficulties in proving that psychological harm took place, challenging the effectiveness of Article 5.1.(a) and its provisions on psychological harms. The Future of Life Institute (FLI 2021:7) explains why it is challenging to prove harm, "subliminal manipulation is hard to detect and because it will be difficult for an affected person to prove a causal relationship between the subliminal manipulation and the harm incurred." Company Mediaset Italia (2021:2) outlines the impossibility of proving the harm as dictated by the nature of subliminal techniques, "a subliminal technique is by its definition not detectable by the person impacted, hence informed consent is not possible, nor is it possible for an individual to prove that his/her/their behaviour was materially distorted." Bits of Freedom (2021:3) also notes this disproportionate difficulty in providing proof of harm, "If passed into law, this will place an

unreasonable burden of proof on individuals to demonstrate future or actual harm, *as it is extremely difficult if not impossible* for individuals to gather information or evidence on these practices." Even a business association, etami, some of whose members are Volkswagen, Zalando, Siemens, and the Technical University of Berlin, raise their concerns surrounding the vagueness of harm provisions, "While the idea is laudable, the wording is vague [...] Especially as seen by the citizen, it may not be clear when one is affected" (etami 2021:2). Despite the presence of the AIA provisions aiming to address psychological harms, the current definition of prohibited uses and the requirement to prove harm significantly limit the scope of prohibited uses. Thus, in the context of manipulation qualities, making it an efficient tool to bypass control and influence behaviours without being noticed, the following impacts discussed are unaddressed by the AIA.

Forms of manipulation. Different manipulative AI impacts are acknowledged throughout investigated feedback submissions despite differing levels of specificity. Bublitz and Douglas (2021:1) define manipulation as, "When AI systems influence thought or behaviour in ways that bypass or weaken rational control, they are manipulative." Women in AI (2021:3) also include exploitation, which they do not define, to the list of negative impacts unaddressed by the AIA, "This provision [Article 5] is insufficient to protect persons in the European Union from other serious harms, such as exploitation." European Evangelical Alliance (EEA 2021:2) outlines the additional impacts of manipulation as "[individual reactions to stimuli] could be used to manipulate, to excessively nudge, or inappropriately interfere with a person's freedom of thought, critical thinking processes, undermining their judgment." Overall, the variations of manipulation practices reduce control over one's behaviour.

There are divergent approaches to the scope of manipulative impacts, one on the individual level and others reaching beyond the individual, thus affecting society at large.

Individual impacts. Some of the individual impacts include the weakening control over one's actions, which is conducted by "technological captivation of attention" (Bublitz and Douglas 2021:6), optimisation used in social media (CHAI 2021:6), and online gaming websites (Bublitz and Douglas 2021:6). The Cambridge Analytica example is used to demonstrate how political targeting can be used to induce impulsive anger across individuals and change their emotional state (Bublitz and Douglas 2021:8). Another mechanism through which manipulative effects can be induced is when human emotional control is bypassed by targeting users in their vulnerable state (Bublitz and Douglas 2021:7). The resulting efficiency of manipulation can be used to alter individual behaviours. Amnesty International (2021:5) outlines the AI effects as nudging users into particular behaviours, "AI has shown an enormous capability to condition people and even manipulate people into certain behaviour, leading to adverse personal, societal or democratic effects."

Social impacts. The FLI outlined how this individual impact could bring societal impacts, "An AI system that maximises ad clicks, for example, will show users addicting content. In turn, this application causes users to spend more time on social platforms and may foster societal polarisation and increased misinformation" (FLI 2021:4). Here, individual impacts are affecting the social sphere. The FLI specifies the impact at the societal level and how it materialises across individual impacts even if they are imperceptible at the individual level:

> AI applications may cause societal-level harms, even when they cause only negligible harms to individuals. For example, a political marketing application may reduce a

> person's desire to vote by a small amount. At an individual level, the impact of this application may not be considered an infringement of fundamental rights, but collectively, the effect may be large enough to change an election result. (FLI 2021:4)

The difference is made between individual harms and societal-level harms, which the FLI (2021:5) describes as "indirect and aggregate harms." As a result, not only individuals are affected by manipulative systems. Society is affected by both the aggregate of individual harms and societal shifts, even if they might be felt less explicitly by individuals.

Human rights protections. Some manipulative influences leading to negative impacts are defined by outlining the need for protections and specifying undermined human rights. Such as exemplified by Bublitz and Douglas (2021:5), "Its [regulation on manipulative influences] absence would leave central aspects of the human person unprotected and fail to respect some of the most important fundamental rights." CHAI (2021:6) proposes to mention Article 3 of the European Union Charter of Fundamental Rights, which acknowledges the right to mental integrity protections in the AIA Explanatory Memorandum. Bublitz and Douglas (2021:1) mention freedom of opinion, mental integrity, and the rights of freedom of thought as being implicated by manipulative AI impacts. As a response to the AI Act prohibiting uses of AI systems inflicting psychological harms, Women in AI (2021:3) proposes to replace the prohibition of uses leading to harms by replacing it with human rights infringements "materially distort[ing] a person's behaviour in a manner that undermines or is likely to undermine the fundamental rights of that person." European Evangelical Alliance proposes to "draft a Declaration of Digital Human Rights," in which the right against being manipulated could be enshrined (2021:4).

4.2.3. Labour Rights and Workers (Topic 12)

The topic's contents are captured by two main subtopics, teachers' and more general workers' rights. The majority of the top 10 entries come from trade unions, out of which two trade unions submitted ETUCE's submission (ČMOS 2021; GEW 2021). As a result, when ETUCE submissions are referenced in subsequent analysis, it represents the position statements of three entities. Seven submissions originate from trade unions, while NGOs submitted two responses and a business organisation submitted one. The granularity of approaches means that the rights discussed are not referred to in the regulation. Therefore, by default, the concerns raised by the submissions are unaddressed by the AIA.

Education and Teachers' Rights

The issues raised by the investigated responses on education are not referring to education being considered a high-risk application area. Instead, the relationship between changes brought by the digitalisation of education and the education profession is of focus. As with other topics investigated with close reading, the impacts brought by AI are outlined in relation to potential human rights infringements, namely the right to equal access to education, as outlined in the Charter of Fundamental Rights of the European Union and the European Pillar of Social Rights (ETUCE 2021:2).

The demonstrations of AI impacts on the reduced quality of education are made on two grounds: a reduced role of teachers and an increase in educational inequalities. From the perspective of

educational personnel as workers and AI impacts on their professional activities, privacy rights and rights to disconnect are also outlined.

The argument for maintaining teachers' autonomy is made in relation to preserving the quality of education, which is approached from a human rights standpoint. ETUCE (2021:2) calls for the prohibition of AI uses "that are designed to replace education personnel or can damage the social value and the quality of education." ETUCE (2021:2) emphasises the role of teachers in education as paramount.

Two sets of risks posed by increased inequality are outlined. On the one hand, education as a public service is challenged by the privatisation of education. The transition to remote learning is used as an example of rising education inequalities, "As the ongoing Corona-pandemic has shown, digitalization has enhanced the inequalities in education (OECD, 2021), so this risk is real, not potential or plausible" (COV 2021:1). The advancements in educational technologies (ed-tech) are referred to as posing risks to education as a public good, "[Increase in AI use in education] poses a real danger for privatization, commercialization and monopolization, all of which endanger the equal access to education" (COV 2021:1-2). This position on risks posed by ed-tech is also maintained by TUI (2021).

On the other hand, educational inequalities can be exacerbated by biases embedded in the AI systems. ETUCE outlines the processes in which biases are created in AI systems due to the lack of diversity across teams creating AI products, "leading to a detrimental impact on inclusion and equality in education" (ETUCE 2021:4). The implementation and operationalisation of AI ethics principles, such as trustworthiness and transparency, are also questioned for their effects on the quality and inclusivity of education (ETUCE 2021:1; TUI 2021).

As for the educational profession, data privacy, intellectual property rights, and the right to disconnect are discussed for their impacts on working conditions (ETUCE 2021:17). The use of digital tools in education exposes users to a range of risks, "AI tools storing a vast amount of data cause inevitable risks on data protection, privacy and intellectual property rights of teachers and academics and other education personnel" (ETUCE 2021:4). The right to disconnect and the protection of teachers' autonomy are called to be maintained (ETUCE 2021:17).

Workers' Rights

Workers' rights are discussed regarding AI use in the workplace. The issues surrounding the absence of the opt-out regime from AI technologies used at work (Negotia 2021:1) and algorithmic management are noted as negative impacts of AI (UNI 2021:1). AI systems are noted for potential discrimination caused by the use of incomplete data for systems training (UNI 2021:2).

This subtopic on labour centres around risks posed by using biometric recognition and categorisation technologies at work. The use of biometric identification and emotional categorisation are noted as harmful surveillance practices (UNI 2021:5). In turn, surveillance poses potential infringements of human rights and safety (UNI 2021:1-2).

Even if only a few documents of topic 12 dealt with broader approaches to workers' rights, the labour market changes brought by AI transformation and the approaches to retraining did not

take a significant role. Only ETUCE (2021) and Eurocities (2021) outlined potentially negative effects of AI and the importance of retraining, "AI […] might cause displacement or even replacement of workers, polarisation of job demand between high-skilled and low-skilled jobs, and worsen the status of already fragile groups of people, such as the digitally excluded, long-term unemployed and low-skilled people" (Eurocities 2021:2). Given the significant role the AI and work discussion takes in popular discourses (Crépel and Cardon 2021) and across national AI strategies, the higher prevalence of AI effects on labour across investigated documents responses could be assumed, which was not the case.

4.2.4. Education and Children (Topic 13)

The importance of combining computational and qualitative approaches is especially apparent when engaging with the top 10 documents on the topic. The discussion of impacts can be induced from the list of most probable words across the topic, yet the extent to which negative impacts are discussed is contextualised with close reading.

The main response framing the impacts on children is submitted by 5Rights Foundation, which also has the highest topic weighting across investigated documents. While other documents have lower probabilities, which was also observed across other topics, the remaining documents do not discuss the negative impacts. Other documents contain words such as young or children, yet the context of their uses is far from discussing negative impacts. For instance, Eurocities discuss the need to increase young women's participation in developing technologies (Eurocities 2021:2), and Thorn (2021) advocates for regulatory flexibility to use technologies, such as facial recognition, to protect children from abuse and exploitation.

The 5Rights response significantly shapes the discussion of the main impacts of AI systems on children. As a result, the following set of impacts is not as conclusive as the impacts outlined by analysing previous topics. In opposition to previous themes, where the agreements and divergence of perceptions on impacts could be inferred across the responses, the impacts represented here are representative of one organisation.

5Rights (2021:8) defines one emerging harm, "compulsive use as an internet harm for children." Persuasive design term is used to capture the phenomena of manipulation, which was discussed in topic 6. Persuasive design occurs when human actions are "accelerated, nudged and determined by technology that, in turn, changes or trains human behaviour" (5Rights 2021:20). Similarly to mechanisms by which rational control over one's behaviour can be reduced, an example of Angry Birds design is used to explain how the design of feedback loops can lead to compulsive behaviours, to which children have increased vulnerability when compared to adults due to their stronger impulse to seek instant gratification (5Rights 2021:20).

In this case of persuasive design used on children, negative manipulation effects are illustrated in greater detail than in topic 6. Persuasive design leads to social anxiety and self-esteem issues because, in persuasive design settings, the focus is placed on quantity, which changes the importance of peer relationships (5Rights 2021:21). Other negative effects supported by referred studies include increased aggression, reduced quality of relationships linked to an increased sense of loneliness linked with long periods of time spent on digital technologies and disagreements with parents (5Rights 2021:29). Loss of sleep and effects on education and memory are referred to as opportunity costs of using technologies embedded in persuasive design methods (5Rights 2021:29-30). The impact of surveillance is not clarified, yet the

potential for social scoring is criticised because it could affect children beyond their developmental stages of experimentation (5Rights 2021:34).

Since only one document deals directly with the negative impacts of AI on children, human rights implications are not as granularly outlined. Nevertheless, it can be assumed that similar human rights risks are posed as outlined in the manipulation topic. Thus, the discussion on human rights is limited to privacy (5Rights 2021:16) and the calls for the creation of new rights, such as the right to conscious use and be informed (5Rights 2021:11).

4.2.5. Affected Human Rights (Topic 16)

As mentioned at the beginning of the guided close reading section, 6 of the 10 most representative responses of topic 16 were discussed in previous sections. To avoid overlap, the remaining 4 responses were attributed to one of the four topics, in which the investigated documents had the highest topic weight. As a result, the potential infringements of human rights had already been discussed in previous sections.

4.3. H1 Discussion

Figure 3 on the next page summarises the different algorithmic impacts discussed in this section. It visualises the distinctions between individual and societal impacts and implicated human rights unaddressed by the AIA. The range of negative social impacts is noted throughout the responses, even if the lines between the categories are not rigid. Individual impacts can accumulate into societal level changes, while societal impacts can affect individuals' lives. The boundary between these impacts is representative of the flexible boundaries set by other studies establishing distinctions between individual and social impacts (Smuha 2021a:4; Vesnic-Alujevic et al. 2020:3). Even if the rigorous discussion could challenge the classification, the boundaries' quality does not determine the quality of the finding that negative impacts are also social.

The negative impacts outlined are social in both conventional conceptualisations of social and more recent distinctions about the emerging types of societal harms resulting from changes brought by AI (Smuha 2021a). Discrimination, surveillance, and the use of social scoring on children affect particular groups, already translating to communal negative impacts. However, the close reading analysis revealed the emergence of negative impacts that concern the political sphere by altering individual behaviours even if at the level unnoticeable to an individual (FLI 2021), at the aggregate level may result in a change in political outcomes.

Figure 3. Summary of main negative AI impacts unmitigated by the AIA

Topic id	Topic label	Individual	Reaching beyond individual (group and social)	Human rights affected	Human rights represented across topic 16
4	Biometric impacts	Discrimination (commercial uses, policing) Psychological chilling effects (individuals adapting their behaviour)	Discrimination (policing) Diminished equality Surveillance	Privacy, autonomy, anonymity, mental integrity, right to opinion and express one's identity, freedom of thought, non-discrimination Chilling effects on human rights (freedom of expression, assembly, movement, protest, and association)	
6	Manipulation	Manipulation of users weakening the control of their own actions (captivation of attention, nudging into certain behaviours, micro-targeting users in their vulnerable states/ altering of their emotional states) Induction of impulsive anger	Altering of political decision-making outcomes (political targeting) Societal polarisation Increased misinformation	Mental integrity Freedom of opinion Right to freedom of thought	
12	Working conditions and education (workers)	Inclusivity, inequalities, and quality of education Teachers' autonomy Algorithmic management, absence of opt-out regimes	Monopolisation of education Surveillance Polarisation of labour market Increased inequalities amongst vulnerable groups	Equal access to education Privacy Intellectual property Workers' rights	
13	Impacts on children	Compulsive use of technologies Persuasive design (manipulation) Social anxiety, aggression, reduced quality of relationships (peer and parents)	Surveillance Social scoring used on children in their experimentation stages	Rights of children Privacy Right to be informed	

From the topics selected for the close reading analysis, some of the negative impacts discussed in topic 6 on manipulation match the qualities of societal harms outlined in the literature review (Slaughter et al. 2021; Smuha 2021a; Tufekci 2015). Societal harms are incremental, gradual, and challenging to note until the resulting impact has already caused significant change (Smuha 2021a).

As discussed previously, the negative impacts of manipulation are hard to notice since they are effective because they bypass rational control and thus are even harder to prove when psychological harms are incurred. When discussing recommender systems, Bublitz and Douglas outline how users of recommender systems come to trust the product, "by cultivating the *gradual development of trust* [emphasis added] in and reliance on recommendations through *repeated interactions* [emphasis added]" (2021:7). This trust can be exploited by occasionally recommending not the most suitable options to the users (Bublitz and Douglas 2021:7). This description of how recommender systems function resonates with the definition of societal harms provided by Smuha (2021a) because, firstly, changes are gradual. Secondly, the manipulative outcomes occur through repeated interactions. Lastly, the changes are hardly noticeable until a considerable impact is induced (Tufekci 2015). According to the qualities of harms outlined in the literature review, the majority of the top 10 documents on the manipulation topic mention impacts whose qualities are associated with this emerging form of impact, which is hardly noticeable and takes shape over repeated interactions, leading to the near impossibility of proving that the impact took place. As outlined, the respondents emphasise the difficulties in measuring and noticing manipulative impacts, which is particularly important in the behavioural field where human control is aimed to be overridden.

The emergence of social impacts is significant because the rights' protections are mainly individual. There is an absence of existing precedents of their enforcement, which is limited to privacy cases. The solutions should either enforce existing rights or create new societal protections.

4.4. H2 Discussion

The close reading analysis reveals three significant trends around levels of certainty surrounding negative AI impacts:

1. There is a widely spread notion across the investigated topics and their responses for a need to increase effort at researching and understanding algorithmic impacts.
2. The impacts of human rights infringements are represented at higher certainty levels. For the rest of the impacts, the communication around certainty and uncertainty can be associated with the length of responses and the level of detail.
3. The topic of biometric recognition systems' impact revealed additional uncertainty of the capabilities of biometric systems, which affects the knowledge base on negative impacts.

The need to further research impacts and risks was outlined across all topics investigated with close reading. In the labour topic, this need was expressed by ETUCE (2021:1), while on the impact on children, 5Rights recommends creating a centre of expertise for investigating impacts on children (5Rights 2021:9). When discussing biometric recognition (topic 4) and its effects on human rights, the need to acquire a more comprehensive understanding of different

categories of privacy, such as general, psychological and identity, is noted as a neglected area that has "received little attention to date" (ALLAI 2021:13). Bublitz and Douglas note the lack of knowledge surrounding the mechanics of a diverse range of manipulation practices, "many new forms of influence are poorly understood, their strength and modes of operation are not always easy to determine" (Bublitz and Douglas 2021:5). Given that the issue of lack of knowledge about impacts is acknowledged across all topics, it can be concluded that more research needs to be dedicated to investigating those impacts to understand how they come into being and how they affect users. As discussed in the literature review, the conflict of interests between corporations and the public good can influence research through corporate funding. As a result, the recommendation by 5Rights to establish an independent research centre for research on AI impacts on children can be extended to AI impacts more generally, where independent departments could focus on either particular technologies used, affected areas or impacted groups, among others.

The level of certainty expressed in discussing human rights infringements contrasts with discussions across the remaining topics. The wealthy scholarship on human rights infringements and their abuses already documented in the privacy and surveillance fields and studies on AI biases can offer one of the explanations for a relative lack of detailed explanations of pathways to human rights infringements. This assumption is exemplified by this Access Now statement:

> "Civil society organisations such as Access Now, Article19 and the AI Now Institute *have long pointed* [emphasis added] to how emotion recognition systems violate a range of human rights, including the right to privacy, right to freedom of expression, right to protest, right against self-incrimination, and the right to equality and non-discrimination." (Access Now 2021:10)

The shortcut from the perspective of human rights infringements is thus taken, as seen in Amnesty International's (2021:3) statement, "We have seen a surge in biometric recognition applications in workplaces, in recruitment and human resources and in commercial settings, leading to corporate surveillance and the widespread use of techniques that constitute a threat to our human rights." This statement reflects the status of the knowledge base around these technologies. It is acknowledged that these technologies are already used for corporate purposes; even if the extent of their use is not specified, the assumption is taken that these uses are prevalent and result in negative impacts, such as violations of human rights.

As a result, the mechanisms according to which human rights infringements occur are not as detailed as when explaining impacts stemming from manipulation or biometric recognition. When behavioural changes, chilling effects or discrimination are discussed, the explanations are supported by actual examples or hypothetical scenarios, which are then explained. Across the manipulation topic, the plausibility surrounding the discussion of manipulative AI impacts can be explained by the difficulties in measuring and proving the changes and the habitual gradual nature of changes. Therefore, a portion of examples discussed in responses contain hypothetical instances. Bublitz and Douglas (2021:7) hypothesise a scenario where an emotionally frustrated social media user is nudged toward content that generates hatred directed against immigrants. This example is used to illustrate how human emotional control can be bypassed by targeting the user in their vulnerable state (Bublitz and Douglas 2021:7). Due to Cambridge Analytica leaks, there is a documented case of such occurrence, which Bublitz and

Douglas refer to following the latter hypothetical example. The level of detail is thus higher even if the hypothetical instance is used, given the lack of publicly available data on how private companies design their ranking algorithms.

Different approaches to whether discussed impacts constitute harm represent another example of varying degrees of certainty and granularity used to make those claims. Across the manipulation topic, several submissions declare the impacts as harms, while others take a more detailed approach to discuss what constitutes harm. Despite not defining psychological harms, manipulation and violations of mental integrity fall under the harms category in the CHAI submission (2021:5-6). Women in AI also link manipulation with harm, "prevent any harm to children and youth by way of algorithmic manipulation" (2021:6). Submissions, where behavioural impacts are discussed briefly, tend to call the impacts harms. On the other hand, the document with the highest topic weight discussed what constitutes harm. Bublitz and Douglas the requirement set by the AIA on psychological harms, which authors explain:

> [T]hey are to be understood as they usually are in law, as physical or psychological injuries, setbacks to health or biological or social functioning. Yet nonconsensual subliminal interventions that significantly alter thought or behaviour are plausibly ethically wrong—because they bypass rational control—even if they inflict no harm. (Bublitz and Douglas 2021:9)

Thus, there are different interpretations of negative AI impacts; some regard them as constituting harms while others treat them as negative effects. Despite considering whether it constitutes harm, these manipulative practices produce negative impacts, which are acknowledged by the top 10 documents with the highest proportion of topic 6.

Lastly, there was widespread agreement on the lack of a scientific base for emotional and biometric categorisation systems (Amnesty International 2021:2-3; Women in AI 2021:1). The CAIDP states that "[c]ategorization systems do not have any scientific validity" (2021:5), while others call it a form of pseudoscience (Access Now 2021:19). The meta-study on facial recognition and emotion identification is referred to as documenting the lack of scientific evidence (Barrett et al. 2019; ALLAI 2021:14; Amnesty International 2021:3). The potential risks of mischaracterisation and discrimination nevertheless lead to real-life impacts (Amnesty International 2021:2). The feedback responses thus raise concerns about biometric categorisation companies' claims of their efficiency (Access Now 2021:10). Thus, the biometric recognition topic reveals that the technological capabilities of technologies can and should be challenged. If the systems do not function as intended or proclaimed, it can lead to negative outcomes. As a result, as falling with the calls by the scientists investigating AI impacts from a long-term perspective, technological capabilities ought to be assessed to understand the impacts better.

5. Conclusion and Policy Recommendations

The AI Act is a significant step toward leveraging opportunities provided by AI while mitigating risks associated with its uses. Nevertheless, informed by the analysis of feedback responses, based on this research, it is recommended not to leave the following AI impacts unmitigated:

- Biometric identification: discrimination, chilling effects on psychology and exercise of human rights, risks of surveillance, and the plethora of human rights infringements, such as the right to privacy, autonomy, anonymity, opinion, and freedom of thought.
- Manipulation of users: captivation of attention, bypassing of rational control, targeting of users at vulnerable states, altering their emotional states as well as societal level impacts such as increased polarisation, misinformation and their effects on political processes.
- Workers and children: digital inequalities, diminished workers' autonomy and societal challenges to public goods such as education, the polarisation of the labour market, and subsequent human rights, such as privacy, children and workers' rights.

The mitigation of these negative impacts could target either the cause or effects level. On the one hand, biometric identification systems and AI systems based on manipulative design could be banned. As noted by feedback responses, there is a lack of evidence supporting the claim that biometric recognition systems are efficient. In addition, manipulation overrides human control, which could be interpreted as immoral, bringing "no obvious social benefit" (ALLAI 2021:6) apart from the economic profits for the companies developing these systems, which could also be then interpreted as providing jobs and contributing to the public good. Nevertheless, the rigid stance prohibiting the use of biometric identification systems and those associated with this technology could be easily expected to receive a strong pushback from the industry stakeholders, even if there is a lack of evidence supporting the efficiency claims of AI systems, such as biometric recognition systems.

On the other hand, a more viable solution would be setting the protections against negative individual and social AI impacts and establishing a redress mechanism to enforce the protections. As demonstrated by this analysis, impacts are not only individual, yet most of the existing protections are individual based. Creating redress mechanisms in the AIA, including mechanisms for redress of individual and collective complaints, is recommended. New rights, such as the right to not be manipulated, and impact specific, such as the right to be informed when an individual is engaging with an AI system or the right to opt out from being treated by an AI system in the work contexts, should be considered. Given that the AIA takes a risk-based approach, it is recommended to consider establishing new legal rights corresponding to new challenges brought by societal AI impacts outside the AIA.

It has to be stressed that this analysis reveals an underlying difficulty in capturing negative AI social impacts since they tend to take place over more extended periods of time through repeated interactions that make them less noticeable, as is the case in manipulative systems. As noted by the feedback respondents, the issue lies in the lack of understanding of these technologies, which translates to the necessity to understand their functioning mechanisms better. The second option puts a relatively hardly carriable burden of proving that the negative impact affected individuals or society. The difficulty in proving hardly noticeable and provable

impacts should not prevent the regulators from creating new rights and protections. As for societal impacts, new institutions, which creation is discussed below, could be established for measuring and monitoring the impacts or extension of responsibilities of existing institutions.

Lastly, to address the commonly noted need expressed in feedback responses to increase our knowledge base around AI impacts, more research is required to disentangle the impacts, the mechanisms of technologies bringing them and monitor their development. As noted by some respondents and identified by the state of knowledge literature identifying informational asymmetries, there is a need for independent monitoring and measuring algorithmic impacts. Whittlestone and Clark (2021) provide directions for which questions to be tackled by governments, arrangements for hybrid approaches to subcontracting that could be realised and pilot projects' ideas corresponding to identified policy challenges. The approach for such a body could be inspired by the European Environmental Agency (Smuha 2021a:17). Thus, it is recommended to establish a body or delegate powers to existing entities and expand the responsibility to develop and enhance measurements of AI impacts and monitor the measurements' implementation.

The impact assessments mainly relying on companies' self-regulation in the AIA could be capitalised on. Following the recommendations set by responsible innovation frameworks promoting participatory technology governance (Stilgoe et al. 2013), the impact assessments of AI systems, potentially influencing social interest, could be made public (Smuha 2021a). Publicly accessible impact assessments would encourage democratic decision-making processes, potentially resulting in social impacts.

The impacts outlined by the case study on unmitigated negative AI impacts are assumed to translate well to challenges other jurisdictions face. This assumption is made based on the technological advancements extending beyond the borders of jurisdictions. For instance, manipulative AI and its mechanisms of overriding rational control are universal to humans; thus, associated negative impacts arising from the use of technologies using manipulation techniques can be presumed to be induced universally. As a result, the mapped impacts are expected to translate to other jurisdictions, such as China or the US, where some examples of negative AI impacts noted by feedback respondents were documented.

5.1. Evaluation of Method and Further Research Directions

The combination of topic modelling with guided closed reading could be regarded as a successful research strategy. Close reading provides a greater level of detail to the impacts investigated, and there is evidently an increased wealth of detail coming from close reading. It is the close reading that provides answers to the research question of what the main negative AI impacts are undressed by regulation. However, TM is an integral part of the research design. Without topics discovered by TM, all of the most significant submissions could not have been analysed, and subsetting would have to be used. As the variance of topic weights of investigated topics demonstrates, sampling in the area of feedback responses could result in the random selection of documents that are not likely to be representative of the question investigated. In this case, TM provided topics and associated words providing context for subset selection for further analysis. The topics discussed because of the method employed were not cherry-picked; they are representative of the entire corpora of feedback responses. However, it has to be noted that impacts of societal importance, such as on the environment, good life and wellbeing, were

not captured by the relatively high number of topics. The model captures the main unmitigated impacts, yet it does not cover a complete mapping of the emerging themes. In addition, the efficiency of selecting the number of topics for the model could be challenged since it relies on the interpretation of what level of granularity of topics is best suited for answering the question. Topics in isolation do not answer the research question; however, combined with guided close reading, they provide guidance on which themes signal those negative AI impacts. As a result, the number of topics is validated by the close reading.

Methodologically, variations of LDA could be further explored to match the topic distributions across the documents better; otherwise, separate TM models could be created for different categories based on document lengths. The potential future directions related to the dataset include investigations of the impact of the feedback responses as the AIA is further shaped and the relationships and power structures revealed by collaboration between different stakeholder groups. This work presents an initial mapping of the negative impacts identified by feedback respondents in response to the AIA.

6. Bibliography

Abdalla, Mohamed, and Moustafa Abdalla. 2021. 'The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity'. Pp. 287–97 in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3461702.3462563.

Ada Lovelace Institute. 2022. 'Algorithmic impact assessment: a case study in healthcare'. https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment- case-study-healthcare/.

Ada Lovelace Institute, and DataKind UK. 2020. 'Examining the Black Box: Tools for Assessing Algorithmic Systems: Identifying Common Language for Algorithm Audits and Impact Assessments'. https://www.adalovelaceinstitute.org/report/examining-the-black- box-tools-for-assessing-algorithmic-systems/.

AIAAIC. n.d. 'AIAAIC Repository'. Accessed 12 March, 2022. https://www.aiaaic.org/aiaaic- repository.

Albu, Oana Brindusa, and Mikkel Flyverbom. 2019. 'Organizational Transparency: Conceptualizations, Conditions, and Consequences'. *Business & Society* 58(2):268–97. doi: 10.1177/0007650316659851.

Alves, Amanda M., Eric Brousseau, Nada Mimouni, and Timothy Yu-Cheong Yeung. 2021. 'Competing for Policy: Lobbying in the EU Wholesale Roaming Regulation'. *Telecommunications Policy* 45(3). doi: 10.1016/j.telpol.2020.102087.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI'. *Information Fusion* 58:82–115. doi: 10.1016/j.inffus.2019.12.012.

Ashurst, Carolyn, Emmie Hine, Paul Sedille, and Alexis Carlier. 2021. 'AI Ethics Statements – Analysis and Lessons Learnt from NeurIPS Broader Impact Statements'. arXiv preprint arXiv:2111.01705v1. doi:10.48550/ARXIV.2111.01705.

Barrett, Lisa Feldman, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. 'Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements'. *Psychological Science in the Public Interest* 20(1):1–68. doi: 10.1177%2F1529100619832930.

Baum, Seth D. 2017. 'A Survey of Artificial General Intelligence Projects for Ethics, Risk, & Policy'. in *GCRI Working Paper 17-1*.doi: 10.2139/ssrn.3070741.

Baum, Seth D. 2020. 'Medium-Term Artificial Intelligence and Society'. *Information* 11(6). doi: 10.3390/info11060290.

Baumer, Eric P. S., David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. 'Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence?' *Journal of the Association for Information Science and Technology* 68(6):1397–1410. doi: 10.1002/asi.23786.

Blei, David M., and John D. Lafferty. n.d. ‘Topic Models’. http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. ‘Latent dirichlet allocation’. *Journal of Machine Learning Research* 3:993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com.

Boussalis, Constantine, and Travis G. Coan. 2016. ‘Text-Mining the Signals of Climate Change Doubt’. *Global Environmental Change* 36:89–100. doi: 10.1016/j.gloenvcha.2015.12.001.

Bradford, Anu. 2020. *The Brussels Effect: How the European Union Rules the World*. New York: Oxford University Press.

Bunnell, Noah Winter. 2021. ‘Remedying Public-Sector Algorithmic Harms: The Case for Local and State Regulation via Independent Agency’. *Columbia Journal of Law and Social Problems* 54(2):261–304. http://blogs2.law.columbia.edu/jlsp/wp-_content/uploads/sites/8/2021/02/Volume-54-Bunnell.pdf.

Campolo, Alex, Madelyn Rose Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. *AI now 2017 report*. AI Now Institute. https://ainowinstitute.org/AI_Now_2017_Report.pdf.

Chui, Michael, Bryce Hall, and Alex Sukharevsky. 2021. ‘The State of AI in 2021’. *McKinsey & Company*. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Global%20survey%20The%20state%20of%20AI%20in%202021/ Global-survey-The-state-of-AI-in-2021.pdf.

Cihon, Peter. 2019. *Technical Report: Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Future of Humanity Institute. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Crépel, Maxime, and Dominique Cardon. 2021. ‘Critique de l’IA dans la presse’. https://medialab.github.io/carnet-algopresse/#/publication/en/.

Cusumano, Michael A., Annabelle Gawer, and David B. Yoffie. 2021. ‘Can Self-Regulation Save Digital Platforms?’ *Industrial and Corporate Change* 30(5):1259–85. doi: 10.1093/icc/dtab052.

Debnath, Ramit, Sarah Darby, Ronita Bardhan, Kamiar Mohaddes, and Minna Sunikka-Blank. 2020. ‘Grounded Reality Meets Machine Learning: A Deep-Narrative Analysis Framework for Energy Policy Research’. *Energy Research & Social Science* 69(101704). doi: 10.1016/j.erss.2020.101704.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. ‘Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding’. *Poetics* 41(6). doi: 10.1016/j.poetic.2013.08.004.

van Dorsser, Cornelis, Warren E. Walker, Poonam Taneja, and Vincent A. W. J. Marchau. 2018. 'Improving the Link between the Futures Field and Policymaking'. *Futures* 104:75–84. doi: 10.1016/j.futures.2018.05.004.

EDPB-EDPS, European Data Protection Board and European Data Protection Supervisor. 2021. 'EDPB-EDPS Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)'. *European Data Protection Board*. https://edpb.europa.eu/our-work-tools/our-_documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-52021-proposal_en.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. 'How to Make Causal Inferences Using Texts'. arXiv preprint arXiv:1802.02163. doi: 10.48550/ARXIV.1802.02163.

Eickhoff, Matthias, and Runhild Wieneke. 2018. 'Understanding Topic Models in Context: A Mixed-Methods Approach to the Meaningful Analysis of Large Document Collections'. in *Pp. 903-912 in Proceedings of the 51st Hawaii International Conference on System Sciences*. http://hdl.handle.net/10125/50000.

European Commission. 2020. 'Artificial intelligence – ethical and legal requirements'. *European Commission*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial- intelligence-ethical-and-legal-requirements_en.

European Commission. 2021a. 'A European Approach to Artificial Intelligence'. *European Commission*. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial- intelligence.

European Commission. 2021b. 'Artificial Intelligence – Ethical and Legal Requirements: Feedback and Statistics: Proposal for a Regulation'. *European Commission*. https://ec.europa.eu/info/law/better- regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal- requirements/feedback_en?p_id=24212003.

European Commission. 2021c. 'Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM/2021/206 Final'. *European Commission*. https://eur- lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206.

European Commission. 2021d. 'Rolling Plan for ICT Standardisation'. *Artificial Intelligence." European Commission*. https://joinup.ec.europa.eu/collection/rolling-plan-ict- standardisation/artificial-intelligence.

Feijóo, Claudio, Youngsun Kwon, Johannes M. Bauer, Erik Bohlin, Bronwyn Howell, Rekha Jain, Petrus Potgieter, Khuong Vu, Jason Whalley, and Jun Xia. 2020. 'Harnessing Artificial Intelligence (AI) to Increase Wellbeing for All: The Case for a New Technology Diplomacy'. *Telecommunications Policy* 44(6). doi: 10.1016/j.telpol.2020.101988.

Felzmann, Heike, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2019. 'Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns'. *Big Data & Society*. doi: 10.1177/2053951719860542.

Floridi, Luciano. 2014. *The 4th Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford, United Kingdom: Oxford University Press.

Floridi, Luciano, and Josh Cowls. 2019. 'A Unified Framework of Five Principles for AI in Society'. *Harvard Data Science Review* 1(1). doi: 10.1162/99608f92.8cd550d1.

Ganesh, Maya Indira, and Emanuel Moss. 2022. 'Resistance and Refusal to Algorithmic Harms: Varieties of "Knowledge Projects."' *Media International Australia* 1–17. doi: 10.1177/1329878X221076288.

Gasser, Urs, Marcello Ienca, James Scheibner, Joanna Sleigh, and Effy Vayena. 2020. 'Digital Tools against COVID-19: Taxonomy, Ethical Challenges, and Navigation Aid'. *The Lancet Digital Health* 2(8). doi: 10.1016/S2589-7500(20)30137-0.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. 'Text as data'. *Journal of Economic Literature* 57(3):535–74. doi: 10.1257/jel.20181020.

Greene, Derek, and James P. Cross. 2017. 'Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach'. *Political Analysis* 25(1):77–94. doi: 10.1017/pan.2016.7.

Griffin, Rachel. 2022. 'Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality'. doi: 10.2139/ssrn.4064738.

Grimmer, Justin, and Brandon M. Stewart. 2013. 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. *Political Analysis* 21(3):267–97. doi: 10.1093/pan/mps028.

Guo, Chonghui, Menglin Lu, and Wei. 2021. 'An improved LDA topic modeling method based on partition for medium and long texts'. *Annals of Data Science* 8(2):331–44. doi: 10.1007/s40745-019-00218-3.

Hadjiyianni, Ioanna. 2020. 'The European Union as a Global Regulatory Power'. *Oxford Journal of Legal Studies* 41(1):243–64. doi: 10.1093/ojls/gqaa042.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. 'Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach'. *The Quarterly Journal of Economics* 133(2):801–70. doi: 10.1093/qje/qjx045.

Hawkins, Amy. 2018. 'Beijing's Big Brother Tech Needs African Faces'. *Foreign Policy*. https://foreignpolicy.com/2018/07/24/beijings-big-brother-tech-needs-african-faces/.

Heidenreich, Tobias, Fabienne Lind, Jakob-Moritz Eberl, and Hajo G. Boomgaarden. 2019. 'Media Framing Dynamics of the "European Refugee Crisis": A Comparative Topic Modelling Approach'. *Journal of Refugee Studies* 32(Special Issue 1). doi: 10.1093/jrs/fez025.

Heikkilä, Melissa. 2022. 'Dutch scandal serves as a warning for Europe over risks of using algorithms'. *Politico*. https://www.politico.eu/article/dutch-scandal-serves-as- a-warning-for-europe-over-risks-of-using-algorithms/.

Ho, Pauline, Kaiping Chen, Anqi Shao, Luye Bao, Angela Ai, Adati Tarfa, Dominique Brossard, Lori Brown, and Markus Brauer. 2021. 'A Mixed Methods Study of Public Perception of Social Distancing: Integrating Qualitative and Computational Analyses

for Text Data'. *Journal of Mixed Methods Research* 15(3):374–97. doi: 10.1177%2F15586898211020862.

Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2021. 'Algorithmic Amplification of Politics on Twitter'. *Proceedings of the National Academy of Sciences* 119(1). doi: 10.1073/pnas.2025334119.

Isoaho, Karoliina, Daria Gritsenko, and Eetu Mäkelä. 2021. 'Topic Modeling and Text Analysis for Qualitative Policy Research'. *Policy Studies Journal* 49(1):300–324. doi: 10.1111/psj.12343.

Jacobs, Thomas, and Robin Tschötschel. 2019. 'Topic Models Meet Discourse Analysis: A Quantitative Tool for a Qualitative Approach'. *International Journal of Social Research Methodology* 22(5):469–85. doi: 10.1080/13645579.2019.1576317.

Jalonen, Harri. 2012. 'The Uncertainty of Innovation: A Systematic Review of the Literature'. *Journal of Management Research* 4(1). doi: 10.5296/jmr.v4i1.1039.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. 'Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey'. *Multimedia Tools and Applications* 78:15169–211. doi: 10.1007/s11042-018-6894-4.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. 'The Global Landscape of AI Ethics Guidelines'. *Nature Machine Intelligence* 1:389–99. doi: 10.1038/s42256-019-0088-2.

Jones, Bryan D., and Frank R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago, IL: University Of Chicago Press.

Kolkman, Daan. 2020. '"F**k the Algorithm"?: What the World Can Learn from the UK's A-Level Grading Fiasco'. *Impact of Social Sciences Blog*. https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/.

Krafft, P. M., Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. 'Defining AI in Policy versus Practice'. in *Pp. 72-78 in AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. doi: 10.1145/3375627.3375835.

Lane, David A., and Robert R. Maxfield. 2005. 'Ontological Uncertainty and Innovation'. *Journal of Evolutionary Economics* 15(1):3–50. doi: 10.1007/s00191-004-0227-7.

Li, Tiffany C. 2021. 'Privacy in Pandemic: Law, Technology, and Public Health in the COVID-19 Crisis'. *Loyola University Chicago Law Journal* 52(3):767–866. doi: 10.2139/ssrn.3690004.

Liu, Hin-Yan, and Matthijs M. Maas. 2021. '"Solving for X?" Towards a Problem-Finding Framework to Ground Long-Term Governance Strategies for Artificial Intelligence'. *Futures* 126. doi: 10.1016/j.futures.2020.102672.

Loukides, Mike. 2021. 'AI Adoption in the Enterprise 2021'.O'Reilly. https://www.oreilly.com/radar/ai-adoption-in-the-enterprise-2021/.

Maas, Matthijs M. 2021. 'Aligning AI Regulation to Sociotechnical Change'. in *Oxford Handbook on AI Governance*, edited by J. Bullock, B. Zhang, Y.-C. Chen, J.

Himmelreich, M. Young, and A. Korinek. and Valerie Hudson. Oxford University Press. doi: 10.2139/ssrn.3871635.

Malik, Hanna Maria, Mika Viljanen, Nea Lepinkäinen, and Anne Alvesalo-Kuusi. 2022. 'Dynamics of Social Harms in an Algorithmic Context'. *International Journal for Crime, Justice and Social Democracy* 11(1):182–95. doi: 10.5204/ijcjsd.2141.

Marjanovic, Olivera, Dubravka Cecez-Kecmanovic, and Richard Vidgen. 2021. 'Algorithmic Pollution: Making the Invisible Visible'. *Journal of Information Technology* 36(4):391–408. doi: 10.1177/02683962211010356.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. 'Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts'. in *Pp. 735-746 in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery. doi: 10.1145/3442188.3445935.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. 'The Ethics of Algorithms: Mapping the Debate'. *Big Data & Society* 3(2). doi: 10.1177/2053951716679679.

Mohamed, Shakir, Marie-Therese Png, and William Isaac. 2020. 'Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence'. *Philosophy & Technology* 33(4):659–84. doi: 10.1007/s13347-020-00405-8.

Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. *Data & Society*. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/.

Nanayakkara, Priyanka, Jessica Hullman, and Nicholas Diakopoulos. 2021. 'Unpacking the Expressed Consequences of AI Research in Broader Impact Statements'. in *Pp. 795-806 n Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York: Association for Computing Machinery. doi: 10.1145/3461702.3462608.

Nelson, Laura K. 2020. 'Computational Grounded Theory: A Methodological Framework'. *Sociological Methods & Research* 49(1):3–42. doi: 10.1177/0049124117729703.

Nemitz, Paul. 2018. 'Constitutional Democracy and Technology in the Age of Artificial Intelligence'. *Philosophical Transactions of the Royal Society A* 376. doi: 10.1098/rsta.2018.0089.

Niklas, Jędrzej, and Lina Dencik. 2021. 'What Rights Matter? Examining the Place of Social Rights in the EU's Artificial Intelligence Policy Debate'. *Internet Policy Review* 10(3). doi: 10.14763/2021.3.1579.

Nordström, Maria. 2021. 'AI under Great Uncertainty: Implications and Decision Strategies for Public Policy'. in *AI & Society:1-12*. doi: 10.1007/s00146-021-01263-4.

Nowacki, Caroline E., Ashby Monk, and Bertrand Decoster. 2021. 'Who Do Sovereign Wealth Funds Say They Are? Using Structural Topic Modeling to Delineate Variegated Capitalism in Their Official Reports'. *Environment and Planning A: Economy and Space* 53(4):828–57. doi: 10.1177/0308518X20951808.

Nowlin, Matthew C. 2016. ‘Modeling Issue Definitions Using Quantitative Text Analysis’. *Policy Studies Journal* 44(3):309–31. doi: 10.1111/psj.12110.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. ‘Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations’. *Science* 366(6464):447–53. doi: 10.1126/science.aax2342.

O’callaghan, Derek, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. ‘An Analysis of the Coherence of Descriptors in Topic Modeling’. *Expert Systems with Applications* 42(13):5645–57. doi: 10.1016/j.eswa.2015.02.055.

Papadopoulos, Theodoros, and Yannis Charalabidis. 2020. ‘What Do Governments Plan in the Field of Artificial Intelligence? Analysing National AI Strategies Using NLP’. in *Pp. 100-111 in Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*. New York: Association for Computing Machinery.doi: 10.1145/3428502.3428514.

Parson, Edward, Alona Fyshe, and Dan Lizotte. 2019. *Artificial Intelligence’s Societal Impacts, Governance, and Ethics: Introduction to the 2019 Summer Institute on AI and Society and Its Rapid Outputs*. AI Pulse. https://aipulse.org/artificial-intelligences-societal-impacts-governance-and-ethics-introduction-to-the-2019-summer-institute-on-ai- and-society-and-its-rapid-outputs/?pdf=527.

Parson, Edward, Richard M. Re, Alicia Solow-Niederman, and Elana Zeide. 2019. *Artificial Intelligence in Strategic Context*. UCLA School of Law, Public Law Research Paper.doi: 10.2139/ssrn.3476384.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. London: Harvard University Press.

Persoonsgegevens. 2021. ‘Tax Administration Fined for Discriminatory and Unlawful Data Processing’. *Autoriteit Persoonsgegevens*. https://autoriteitpersoonsgegevens.nl/en/news/tax-administration-fined-discriminatory-and-unlawful-data-processing.

Prunkl, Carina E. A., Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. ‘Institutionalizing Ethics in AI through Broader Impact Requirements’. *Nature Machine Intelligence* 3:104–10. doi: 10.1038/s42256-021-00298-y.

Prunkl, Carina, and Jess Whittlestone. 2020. ‘Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society’. Pp. 138–43 in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3375627.3375803.

Quittkat, Christine. 2011. ‘The European Commission’s Online Consultations: A Success Story?: The European Commission’s Online Consultations’. *JCMS: Journal of Common Market Studies* 49(3):653–74. doi: 10.1111/j.1468-5965.2010.02147.x.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. ‘Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing’. Pp. 33–34 in *Proceedings of the 2020 Conference on Fairness,*

Accountability, and Transparency (FAT '20*. New York: Association for Computing Machinery. doi: 10.1145/3351095.3372873.

Rathje, Steve, Jay J. Bavel, and Sander Linden. 2021. 'Out-Group Animosity Drives Engagement on Social Media'. *Proceedings of the National Academy of Sciences* 118(26). doi: 10.1073/pnas.2024292118.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. AI Now Institute. https://ainowinstitute.org/aiareport2018.pdf.

Rhue, Lauren. 2018. 'Racial Influence on Automated Perceptions of Emotions'. doi: 10.2139/ssrn.3281765.

Rhue, Lauren. 2019. 'Emotion-Reading Tech Fails the Racial Bias Test'. *The Conversation*. https://theconversation.com/emotion-reading-tech-fails-the-racial-bias-test- 108404.

Rikap, Cecilia, and Bengt-Åke Lundvall. 2020. 'Big Tech, Knowledge Predation and the Implications for Development'. *Innovation and Development:1-28*. doi: 10.1080/2157930X.2020.1855825.

Roberts, Huw, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2021. 'The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation'. *AI & Society* 36:59–77. doi: 10.1007/s00146-020-00992-2.

Rodrigues, Rowena. 2020. 'Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities'. *Journal of Responsible Technology* 4(100005). doi: 10.1016/j.jrt.2020.100005.

Rosca, Constanta, Bogdan Covrig, Catalina Goanta, Gijs Dijck, and Gerasimos Spanakis. 2020. 'Return of the AI: An Analysis of Legal Research on Artificial Intelligence Using Topic Modeling'. Pp. 3-10 in NLLP@ KDD. http://ceur-ws.org/Vol-2645/paper1.pdf.

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. 'What's Next for AI Ethics, Policy, and Governance? A Global Overview'. in *Pp. 153-158 in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York: Association for Computing Machinery. doi: 10.1145/3375627.3375804.

Schofield, Alexandra, Måns Magnusson, and David Mimno. 2017. 'Pulling out the Stops: Rethinking Stopword Removal for Topic Models'. in *Pp. 432-436 in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia: Association for Computational Linguistics. https://aclanthology.org/E17-2069.

Schradie, Jen. 2019. *The Revolution That Wasn't: How Digital Activism Favors Conservatives*. Cambridge: Harvard University Press.

Schwab, Klaus. 2016. *The Fourth Industrial Revolution*. Geneva: World Economic Forum.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. 'Fairness and Abstraction in Sociotechnical Systems'. P. 3287598 in

Pp. 59–68 in Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT '19*. Atlanta: ACM Press. doi: 10.1145/3287560.3287598.

Slaughter, Rebecca Kelley, Janice Kopec, and Mohamed Batal. 2021. 'Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission'. *Yale Journal of Law & Technology* 23(1). https://yjolt.org/sites/default/files/23_yale_j.l._tech._special_issue_1.pdf.

Smuha, Nathalie A. 2021a. 'Beyond the Individual: Governing AI's Societal Harm'. *Internet Policy Review* 10(3). doi: 10.14763/2021.3.1574.

Smuha, Nathalie A. 2021b. 'From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence'. *Law, Innovation and Technology* 13(1):57–84. doi: 10.1080/17579961.2021.1898300.

Stahl, Bernd Carsten, Job Timmermans, and Catherine Flick. 2017. 'Ethics of Emerging Information and Communication Technologies: On the implementation of responsible research and innovation'. *Science and Public Policy* 44(3):369–81. doi: 10.1093/scipol/scw069.

Stanford Institute for Human-Centered Artificial Intelligence. 2021. 'Artificial Intelligence Index Report 2021'. https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index- Report_Master.pdf.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. 2013. 'Developing a Framework for Responsible Innovation'. *Research Policy* 42(9):1568–80. doi: 10.1016/j.respol.2013.05.008.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019. 'Technology, Autonomy, and Manipulation'. *Internet Policy Review* 8(2). doi: 10.14763/2019.2.1410.

Thomas, Rachel, and David Uminsky. 2020. 'The Problem with Metrics is a Fundamental Problem for AI'. arXiv preprint arXiv:2002.08512. doi: 10.48550/ARXIV .2002.08512.

Tufekci, Zeynep. 2015. 'Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency'. *Colorado Technology Law Journal* 13(2):203–18. https://ctlj.colorado.edu/wp-content/uploads/2015/08/Tufekci-final.pdf.

Ulnicane, Inga, Damian Okaibedi Eke, William Knight, George Ogoh, and Bernd Carsten Stahl. 2021. 'Good Governance as a Response to Discontents? Déjà vu, or Lessons for AI from Other Emerging Technologies'. *Interdisciplinary Science Reviews* 46(1–2):71–93. doi: 10.1080/03080188.2020.1840220.

Ulnicane, Inga, William Knight, Tonii Leach, Bernd Carsten Stahl, and Winter-Gladys Wanjiku. 2021. 'Framing Governance for a Contested Emerging Technology: Insights from AI Policy'. *Policy and Society* 40(2):158–77. doi: 10.1080/14494035.2020.1855800.

Veale, Michael, and Frederik Zuiderveen Borgesius. 2021. 'Demystifying the Draft EU Artificial Intelligence Act'. *SocArXiv*. doi: 10.31235/osf.io/38p5f.

Vervloesem, Koen. 2020. ‘How Dutch Activists Got an Invasive Fraud Detection Algorithm Banned’. *Algorithm Watch*.6 April. https://algorithmwatch.org/en/syri-netherlands-algorithm/.

Vesnic-Alujevic, Lucia, Susana Nascimento, and Alexandre Pólvora. 2020. ‘Societal and Ethical Impacts of Artificial Intelligence: Critical Notes on European Policy Frameworks’. *Telecommunications Policy* 44(6). doi: 10.1016/j.telpol.2020.101961.

Vetulani-Cęgiel, Agnieszka, and Trisha Meyer. 2021. ‘Power to the People? Evaluating the European Commission’s Engagement Efforts in EU Copyright Policy’. *Journal of European Integration* 43(8):1025–43. doi: 10.1080/07036337.2020.1823382.

Wang, Pei. 2019. ‘On Defining Artificial Intelligence’. *Journal of Artificial General Intelligence* 10(2):1–37. doi: 10.2478/jagi-2019-0002.

Wells, Georgia, Jeff Horwitz, and Deepa Seetharaman. 2021. ‘Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show’. *Wall Street Journal*.14 September 2021. https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company- documents-show-11631620739.

Whittaker, Meredith, Meryl Alper, Cynthia L. Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, and Marcel Salas. 2019. ‘Disability, Bias, and AI’. *AI Now Institute*.https://ainowinstitute.org/disabilitybiasai-2019.pdf.

Whittlestone, Jess, and Jack Clark. 2021. ‘Why and How Governments Should Monitor AI Development’. arXiv preprint arXiv:2108.12427. doi: 10.48550/ARXIV.2108.12427.

Winner, Langdon. 1980. ‘Do artifacts have politics?’ *Daedalus* 109(1):121–36. https://www.jstor.org/stable/20024652.

Yam, Josephine, and Joshua August Skorburg. 2021. ‘From Human Resources to Human Rights: Impact Assessments for Hiring Algorithms’. *Ethics and Information Technology* 23:611–23. doi: 10.1007/s10676-021-09599-7.

Yan, Xiaohui, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. 2013. ‘Learning Topics in Short Texts by Non-Negative Matrix Factorization on Term Correlation Matrix’. in *Pp. 749-757 in Proceedings of the 2013 SIAM International Conference on Data Mining*.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

7. Data Appendix 1. Datasets for Guided Close Reading

Each topic is constituted from the top 10 feedback responses ranked according to the weights of topic probabilities.

Figure 1. Impacts of biometric technologies' use (topic 4)

Doc rank	Document name	Name of respondent	Name used for citation	User type	Topic probability	Link to the response
1	090166e5e09dc8f3.pdf	Access Now	Access Now	NGO	0.599	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665462_en
2	090166e5e0aab4b5.pdf	Avaaz Foundation	Avaaz	NGO	0.478	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665625_en
3	090166e5dfee2f58.pdf	Mireille Hildebrandt, Vrije Universiteit Brussel	Hildebrandt	Other	0.371	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2662611_en
4	090166e5e04fc953.pdf	Center for AI & Digital Policy (CAIDP)	CAIDP	Academic/ research Institution	0.264	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2663310_en
5	090166e5e0ab1f08.pdf	Amnesty International	Amnesty International	NGO	0.252	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665634_en
6	090166e5e093520b.pdf	Centre for Commercial Law, School of Law, University of Aberdeen	CCLUA	Academic/ research Institution	0.249	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665397_en
7	Feedback from: European Disability Forum (EDF).pdf	European Disability Forum (EDF)	EDF	NGO	0.234	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2663268_en
8	090166e5e0ab0338.pdf	ALLAI	ALLAI	NGO	0.232	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665629_en
9	090166e5e0ab8064.pdf	EDRI European Digital Rights	EDRi	NGO	0.229	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665234_en
10	090166e5e0a88591.pdf	Women in AI Austria, Carina Zehetmaier	Women in AI	NGO	0.227	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665578_en

Figure 2. Manipulation (topic 6)

Doc rank	Document name	Name of respondent	In-text citation	User type	Topic probability	Link to the response
1	090166e5e0ab49da.pdf	Dr. Jan Christoph Bublitz, Prof. Thomas Douglas	Bublitz and Douglas	Other	0.691	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665640_en
2	090166e5e0aba041.pdf	UC Berkeley Center for Human-Compatible AI	CHAI	Academic/ research Institution	0.196	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665648_en
3	090166e5e0a88591.pdf	Women in AI Austria	Women in AI	NGO	0.173	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665578_en
4	090166e5e0a897ee.pdf	European Evangelical Alliance	EEA	NGO	0.167	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665580_en
5	Feedback from The Value Engineers.pdf	The Value Engineers	Value Engineers	Company/ business organisation	0.148	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2324448_en
6	090166e5df44720d.pdf	Kaspar Rosager Ludvigsen	Ludvigsen	EU citizen	0.138	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2660610_en
7	090166e5e0a6d2f2.pdf	Future of Life Institute	FLI	NGO	0.099	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665546_en
8	090166e5e09d6fcc.pdf	Bits of Freedom	Bits of Freedom	NGO	0.086	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665458_en
9	090166e5e0665fe9.pdf	etami	etami	Business association	0.083	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665168_en
10	090166e5e09b74c5.pdf	Mediaset Italia S.p.A.	Mediaset	Company/ business organisation	0.081	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665444_en

Figure 3. Work conditions and workers' rights (topic 12)

Doc rank	Document name	Name of respondent	Name used for in-text citation	User type	Topic probability	Link to the response
1	090166e5e0743e13.pdf	German Education Union (GEW)*	GEW	Trade union	0.845	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665205_en
2	090166e5e00ea1a9.pdf	ČMOS PŠ*	ČMOS	Trade union	0.837	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2662780_en
3	090166e5e06237cd.pdf	ETUCE*	ETUCE	Trade union	0.744	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2663486_en
4	Feedback from: Teachers' Union of Ireland.pdf	Teachers' Union of Ireland	TUI	Trade union	0.613	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2661971_en
5	090166e5e083274c.pdf	COV (Christelijk Onderwijzersverbond)	COV	Trade union	0.474	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665252_en
6	090166e5de8d4aa7.pdf	UNI Europa	UNI	Trade union	0.365	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2636017_en
7	090166e5e0a4b122.pdf	Negotia	Negotia	Trade union	0.352	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665518_en
8	090166e5e0ab240a.pdf	World Employment Confederation - Europe	WEC	Company/ business organisation	0.338	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665638_en
9	090166e5e03d67db.pdf	Eurocities	Eurocities	NGO	0.297	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2663127_en
10	090166e5e0a6ef1e.pdf	European Edtech Alliance	EEdA	NGO	0.279	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665550_en

*Two organisations (GEW and ČMOS) submitted ETUCE's response, meaning that identical document was repeated for three times in the dataset.

Figure 4. Children and impacts of their use of technologies (topic 13)

Doc rank	Document name	Name of respondent	Name used in-text for citation	User type	Topic probability	Link to the response
1	090166e5e08678db.pdf	5Rights Foundation	5Rights	NGO	0.696	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665266_en
2	090166e5e0884838.pdf	Thorn	Thorn	NGO	0.212	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665284_en
3	090166e5de3652b5.pdf	SAZKA Group a.s.	SAZKA	Company/business organisation	0.122	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2488672_en
4	090166e5e1cd163c.pdf	University of Central Lancashire Cyprus campus	UCLC	Academic/research Institution	0.115	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665299_en
5	090166e5de7d637e.pdf	Artificial and Natural Intelligence Toulouse Institute (ANITI)	ANITI	Academic/research Institution	0.090	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2635975_en
6	090166e5e0a897ee.pdf	European Evangelical Alliance	EEA	NGO	0.082	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665580_en
7	090166e5e09d0637.pdf	BMW	BMW	Company/business organisation	0.080	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665452_en
8	090166e5e0aa234b.pdf	The Future Society	TFS	NGO	0.072	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665611_en
9	090166e5e03d67db.pdf	Eurocities	Eurocities	NGO	0.071	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2663127_en
10	090166e5e09e3011.pdf	Arthur's Legal, Strategies & Systems	Arthur’s Legal	Company/business organisation	0.064	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665467_en

Table 5. Human rights (topic 16)

Doc rank	Document name	Topic in which it was analysed	Name of respondent	Name used for in-text citation	User type	Topic 16 probability	Second most probable topic	Link to the response
1	090166e5e0ab8064.pdf	**4**	EDRI European Digital Rights (EDRi)	EDRi	NGO	0.640		https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665234_en
2	090166e5e080c7c7.pdf	**4 (EDRi's submission)**	Digitalcourage e.V.	Digital Courage	NGO	0.635		https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665649_en
3	090166e5e0ab1f08.pdf	**4**	Amnesty International	Amnesty International	NGO	0.243		https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665634_en
4	090166e5e031b144.pdf		European Center for Not-for-Profit Law (ECNL)	ECNL	NGO	0.220	4: 0.164	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2663061_en
5	090166e5e0a897ee.pdf	**13**	European Evangelical Alliance	EEA	NGO	0.187		https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665580_en
6	090166e5e0ab0338.pdf	**4**	ALLAI	ALLAI	NGO	0.164		https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665629_en
7	090166e5e093f784.pdf		AW AlgorithmWatch gGmbH	Algorithm Watch	NGO	0.152	4: 0.066	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665406_en
8	090166e5dfb6428b.pdf		Civil Liberties Union for Europe	Liberties	NGO	0.149	4: 0.149	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2662292_en
9	090166e5e0a170d6.pdf		The Electronic Privacy Information Center	EPIC	Academic/research Institution	0.139	4: 0.121	https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665484_en
10	090166e5e09d6fcc.pdf	**6**	Bits of Freedom	Bits of Freedom	NGO	0.127		https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/details/F2665458_en

Public Policy Master's Thesis Series

This series presents the Master's theses in Public Policy and in European Affairs of the Sciences Po School of Public Affairs. It aims to promote high-standard research master's theses, relying on interdisciplinary analyses and leading to evidence-based policy recommendations.

Unregulated Negative Impacts of AI:

Mixed Methods Analysis of Feedback Responses to the EU AI Act Proposal

Martyna Kalvaitytė

Abstract

Negative AI impacts are increasingly more noticeable, presenting regulators with the challenge of balancing the opportunities and risks associated with AI. The AI Act Proposal of the European Commission undertakes this challenging task. Two hundred sixty-six feedback responses to the Proposal are analysed using a proposed mixed method to tackle the question of what are the main negative impacts of AI that regulators have failed to address. The study contributes to the literature on adverse AI impacts by offering a mapping of the cross-sectoral impacts and noting their different qualities. Through topic modelling, it is found that the main negative AI impacts are centred around manipulation, the use of biometric recognition systems, adverse effects on workers and children's groups, and overarching potential human rights violations. Guided close reading of identified impact groups' most representative feedback responses illustrates that impacts are both individual and social, emphasising the issue of the lack of protections against societal level impacts. Close reading also provides a use case of algorithmic impacts' descriptions, exemplifying qualities of negative AI impacts outlined by Smuha (2021a) and Tufekci (2015). It is recommended to address the identified individual and social effects by creating protections against societal impacts and establishing redress mechanisms for claiming individual, communal and social remedies. Following identified agreement across investigated responses, it is recommended to establish an independent institution tasked with measuring and monitoring AI systems to increase the knowledge base surrounding the extent of negative AI impacts and the mechanics in which they come into being.

Key words

Artificial intelligence, AI governance, regulation, European Union, societal AI impacts, manipulation