Post-GPT Policy: Risk and Regulation in EU AI Discourse

Kaisla Kajava[1], Emily Öhman[2], Nicole Miu Takagi[2], Elissa Nakajima-Wickham[2], Fedor Vitiugin[1]

[1]Aalto University, Finland
[2]Waseda University, Japan
kaisla.kajava@aalto.fi, ohman@waseda.jp

Abstract

The ubiquity of artificial intelligence (AI) systems across society has generated discussion around risks posed by general-purpose AI (GPAI) technologies. At the same time, harmonized AI regulation is being devised in the European Union (EU) to ensure responsible AI development and uptake. This paper presents a mixed-methods study combining approaches from natural language processing (NLP) and qualitative content analysis to study shifting risk discourse in EU policy-related documents and media articles. The research is anchored around two recent events in the AI space: the popularization of GPAI following the release of ChatGPT and the introduction of the EU AI Act (AIA) regulation. The findings reveal a discursive emphasis on cybersecurity, regional imbalance of computational resources, education and AI literacy, and the implications of risks occurring at scale.

Introduction

With the rapid adoption of artificial intelligence (AI) systems throughout society, researchers, regulators, and industry stakeholders have assessed the potential implications and risks of AI technologies across sectors (Floridi et al. 2021; Yurrita et al. 2022). The escalation of this discourse parallels the increasing adoption of general-purpose AI (GPAI), fueled especially by the launch of ChatGPT at the end of 2022, trained on large-scale data and applicable to a variety of downstream tasks. While efforts in the European Union (EU) striving for harmonized regulation, standardization, and guidelines for the responsible development of AI are being introduced, the emerging risks of AI systems are continuously negotiated and reframed. Moreover, the way AI technologies are presented to and by policymakers, technology experts, and science journalists, among other actors, substantiates the technologies' qualities in both public and policy perception (Johnson and Verdicchio 2017; Cave et al. 2018; Elish and Boyd 2018). AI policy discussions, therefore, offer a perspective into the construction of narratives by policymakers, policy communicators, and diverse other stakeholders, which influences wider AI reception.

Our primary focus in this paper is to examine how the sudden proliferation of GPAI has affected the discourse surrounding EU AI policy in terms of AI-based risks. We collect online news articles, blog and forum posts, and documents published by organizations engaging with EU AI policy from which we generate topics of interest using topic modeling and qualitative content analysis. We use these results to analyze the narrative shifts that have taken place both before and after the launch of ChatGPT in late November 2022. Our data is focused on the time between January 2021 and September 2023, starting several months before the release of the first version of the European AI Act (AIA) proposal in April 2021 (European Commission 2021) until close to a year following the release of ChatGPT. The data is therefore situated around the transition of what has come to be popularly named the "post-GPT" era, before the finalization or applicability of AI-specific regulation.

Employing a mixed-methods approach combining natural language processing (NLP) and qualitative analysis allows us to more comprehensively study risk discourse in EU policy-related documents and policy journalism anchored around recent events in the AI space. We draw insights from the theory of securitization (Balzacq 2011) to discuss the language of risk around AI technologies in the EU as a device to promote the region's competitiveness in the digital economy and to legitimize its position as a global regulatory actor (Kassim et al. 2013; Bradford 2020). We reflect on emerging AI policy discourse around risks newly emerg-

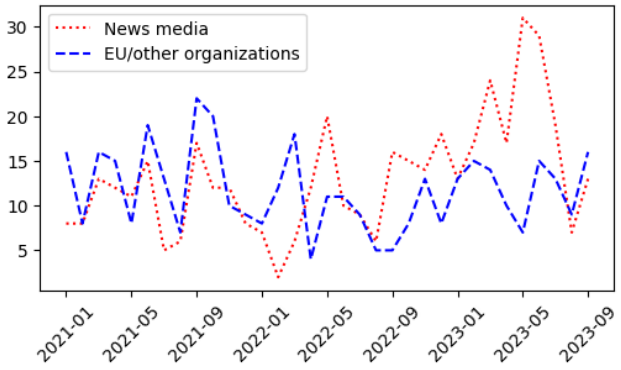


Figure 1: The monthly distribution of EU-related documents and news articles in the corpus published between January 2021 and September 2023.

ing from GPAI as highlighted in previous work and in our empirical findings.

As previous work has shown that language steers AI policy (Ulnicane et al. 2021, 2022), we hold that the significant attention to GPAI applications along with their accessible user experience influences public, expert, and policy perspectives on what AI is, what it can do, what implications it has, and what should be done about it. Despite the fast-paced progress of GPAI affecting all layers of society, policy discourses pre- and "post-GPT", which influence perceptions of risks and regulatory processes, have received limited academic attention. Furthermore, the framing of the discussion is constantly evolving with stakeholders having to adapt their positions to new developments in GPAI, the technology industry at large, as well as regulations.

We start by presenting previous work and providing detailed background information on the AIA and GPAI, followed by introducing our data and methods. We then present the results of both the quantitative and qualitative analyses situating the results in a broader discussion of AI risks as they relate to EU policy discourse and the AIA. We conclude the paper with a discussion of limitations and ethical considerations.

Background

AI in the EU: Risk, Regulation, and Discourse

In April 2021, the European Commission (EC) announced the AI Act (AIA) as a proposed regulatory framework specific to AI systems. As of December 2023, the Council of the European Union (CE) and the European Parliament (EP) have reached a provisional agreement on the AIA, expected to become fully applicable by 2026. The AIA is a product of the EU's regulatory agenda, as part of their digital strategy to make Europe "fit for a digital age"[1]. The regulation posits that AI has the potential to impact individuals and society on a profound scale, and aims to strike a balance between promoting innovation and safeguarding fundamental rights. It uses a risk-based categorization of AI systems to determine their level of requirements and obligations. AI systems that are classified as high-risk, such as those used in essential public services or healthcare, are subjected to stricter rules to mitigate AI-driven harm.

In 2019, the EC High-Level Expert Group on AI (AI HLEG) released a set of ethical guidelines for trustworthy AI (High-Level Expert Group on AI 2019). The AI HLEG guidelines and the subsequent 2020 Assessment List for Trustworthy Artificial Intelligence (ALTAI), provide seven non-exhaustive requirements for the "development, deployment, and use" of trustworthy AI: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination, and fairness, 6) environmental and societal well-being, and 7) accountability.

The requirements reflect various underlying risks in AI development. For example, lack of human oversight may lead to less reliable decision-making, insufficient robustness could place a system at risk for cyberattacks, inadequate privacy measures could put users' personal data at risk, and so on. It is worth noting that the requirements may be complex to implement in practice, as exemplified by Laux (2023) in the context of human oversight.

Rapid AI uptake and complex multi-stakeholder interdependencies have made governing AI reactive and competitive (Smuha 2021; Ulnicane 2022). The EC has set a high priority on AI strategy and sought to secure the EU's position as an influential global actor in regulation (Smuha 2021; Bradford 2020, 2023) and standard setting (Veale and Zuiderveen Borgesius 2021). To legitimize the need for regulation, the adverse effects of a lack of regulation need to be recognized. In this sense, a delay in the "race to AI regulation" (Smuha 2021) can be examined as a security issue both from the standpoint of EU competitiveness and the advancement of societally significant innovation as well as the technology-specific risks pertinent to the deployment of AI technologies.

In the global context, the AIA is part of the EU's efforts to securitize digital technologies through regulation (Mügge 2023). Unlike traditional security concerns, the borderless realm of global technologies may force states to strategize differently, especially as digital technologies primarily remain in the hands of private actors (Mügge 2023). The EU is not alone in this process, as the United States has released its own AI Bill of Rights[2] and Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[3], while China has released its Interim Administrative Measures for Generative Artificial Intelligence Services [4].

It then becomes helpful to understand the process of securitization as it relates to language and the framing of risks or threats. Securitization theory holds that governments construct threats around particular issues through language, which plays a central role in framing and shaping citizens' realities (Eroukhmanoff 2017). A "speech act" is thus a "securitization move" by a government to begin to create a reality where the issue is a threat to a referent object through the use of language. Full securitization occurs when citizens accept this framing as legitimate (Balzacq 2011). Security threats, therefore, under securitization theory, are more contextual than concrete, shifting over time and adapting to specific political environments.

GPAI: Definitions, Risks, and Challenges

In this paper, we use the term general-purpose AI (GPAI) (as opposed to GPAI system) to denote both the underlying pre-trained models and the systems implementing them. GPAI models and systems may or may not have been intended or designed for specific downstream tasks but have broad applicability across many tasks, either out-of-the-box or via

[1]https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age_en

[2]https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[3]https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[4]http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

fine-tuning. Common to these models, in addition to generalizability, is the significant computational resources and data used for training: GPT-4 was reportedly trained on 300 billion words (570 GB of data), whereas BERT was trained on "only" 3.3 billion words (Devlin et al. 2018; Achiam et al. 2023). Many GPAI models are open source and deployable in both commercial and non-commercial systems by individuals or organizations. A key regulatory challenge is allocating responsibility between developers and downstream providers (Engler and Renda 2022).

There exist several resources of AI risks. The recently released MIT AI Risk Repository organizes more than 700 AI-related risks into categories based on their evidenced cause and domain (Slattery et al. 2024). The repository of AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC)[5] contains entries of various controversial events or incidents reported by news outlets. The AI Incident Database (AIID) (McGregor 2021) collects crowdsourced reports of AI-related incidents and also includes taxonomies, such as the CSET AI Harm Taxonomy (Hoffman et al. 2023). AIRO is a hierarchical ontology of AI risk concepts designed based on the AI Act and ISO 31000 risk management standards (Golpayegani, Pandit, and Lewis 2022). However, these resources and taxonomies cover any type of AI system. In this section, we overview two proposed taxonomies of risks specific to LLMs and text-to-image (TTI) models, such as Stable Diffusion, Midjourney, and DALL·E.

Weidinger et al. (2022) present an overview of risks related to LLMs. Their taxonomy divides the risks into six areas: (1) discrimination, hate speech, and exclusion, (2) information hazards, (3) misinformation harms, (4) malicious uses, (5) human-computer interaction harms, and (6) environmental and socioeconomic harms. Many of these areas correspond to the 2019 AI HLEG requirements for trustworthy AI.

In the categorization, human-computer interaction harms caused by LLMs are situations where "humans are deceived or made vulnerable via direct interaction with a powerful conversational agent". This can occur due to unwarranted levels of trust or reliance caused by a conversational system's anthropomorphic characteristics and a tendency to reveal private information in conversation. These issues may provide opportunities for user nudging and manipulation.

Bird, Ungless, and Kasirzadeh (2023) propose a typology of risks specific to generative TTI models divided into three main categories of (1) discrimination and exclusion, (2) misuse, and (3) mis- and disinformation. Discrimination and exclusion encompass social biases such as the amplification of racist, sexist, or ageist structures by algorithmic mediation, as well as job loss for creatives whose labor may be outsourced to TTI models. Misuse covers risks involving sexual, violent, or taboo imagery along with issues of privacy, copyright, and cybersecurity. Privacy and copyright issues stem from models reproducing training data verbatim or near-verbatim, while cybersecurity risks involve adversarial manipulation mainly by malicious actors. The final category of mis- and disinformation includes risks such as the spread of harmful content, polarization, and threats to democratic processes and socio-political stability.

[5]https://www.aiaaic.org/aiaaic-repository

Data

Policy discourse occurs across many genres and document types (af Malmborg 2023). The data consists of online documents from sources affiliated with or covering EU AI policy between January 1st 2021 and September 30th 2023. We focus on this period to cover the emergence of both chatGPT and the first proposal of the EU AI Act. Table 1 presents an overview of the data sources and the types and number of documents collected from each source.

We collected data from organizations focused on or engaged with AI in the EU that have published textual documents online during the analyzed period. We also collected news articles from policy-focused media outlets and prominent blog and forum posts. While many of the organizations listed are independent, they collaborate with or are supported by EU member states or bodies, thus contributing to the discursive construction of EU AI policy. Public blogs, forums, news, policy reports, and organizational announcements all play a role in legitimizing policy efforts.

We selected Euractiv, EUobserver, and Politico Europe for their consistent coverage of EU policy. News media were included due to their fast and reactive publication cycles, with increased AI-related coverage serving as an indicator of the topic's rising policy relevance and discursive momentum. Figure 1 shows the monthly publication frequency of documents and articles across the analyzed period.

We used the keywords *AI* and *artificial intelligence* to find relevant documents in each data source based on tags or by searching the document content. The metadata for each document includes the title, main text, publication date, language, and URL. The language of all documents is English. Altogether, the dataset includes 806 unique documents, of which 432 are news media articles and 374 other types of documents, such as blog posts, news posts by organizations, and reports published by the EC's knowledge service. Although all data sources have some degree of relation to the EU or EU policy-making, the AI Alliance and AI4EU also enmesh industry voices given their multi-stakeholder makeup. The data does not include the legal text of the AIA regulation nor the accompanying impact assessment. However, in contextualized discussion of the data and findings, we take into account those documents along with the general policy climate outlined in EC communications.

This study uses data in the English language due to its predominance in EU-wide discussions and its use as a lingua franca in international and multi-stakeholder fora. While a multilingual analysis would provide a nuanced perspective on local discourses, being especially valuable for citizens' discussions, it involves practical challenges, such as the need for sophisticated translation pipelines and potential variations in semantic interpretations of risk-related concepts, which reach beyond the scope of this study. The use of English-language data allows for a focused analysis of overarching trends in a shared discursive space.

While no dataset can fully capture discourse, the collected data provides a well-rounded overview of AI policy discus-

Name	Organization type	Document type	#
AI4EU	Multi-stakeholder consortium	News posts	68
European AI Alliance	Multi-stakeholder forum	Blog and forum posts	183
Joint Research Centre (JRC)	Knowledge service	Policy reports and briefs	23
Centre for European Policy Studies (CEPS)	Research organization	Publication news	13
European Laboratory for Learning and Intelligent Systems (ELLIS)	Research network	News and event posts	87
EUobserver	News organization	News articles	72
Politico Europe	News organization	News articles	74
Euractiv	News organization	News articles	286

Table 1: Data sources and document types per source.

sion in the EU context. It excludes civic, non-organizational discourse, which falls outside the scope of this study. Moreover, as we draw empirical insights from original textual documents, we do not analyze, e.g., transcripts of public policy debates within European co-legislative bodies. These forms of data remain, however, promising for future research.

A point of consideration is that the analysis does not aim to follow the development of specific policies tackling GPAI or stakeholder perspectives on the legal text of the AIA, but rather to examine broader discursive trends and shifts occurring in policy-related deliberations around AI risks. While the publication of policy reports may lag behind rapid technological developments, blogs and news sources have a quicker reaction time. The study, therefore, treats the release of ChatGPT as a temporal catalyst for shifting GPAI discussions.

Methods

We conducted a time series analysis utilizing guided topic models (Li et al. 2018) to obtain a comprehensive understanding of temporal shifts in topics within our data.

Some argue that different types of topic models, such as BERTopic and Latent Dirichlet Allocation (LDA), have different strengths and weaknesses and suggest using several approaches to extract relevant, consistent, and coherent insights (Axelborn and Berggren 2023; Albalawi, Yeap, and Benyoucef 2020; Egger and Yu 2021). While BERTopic has demonstrated superior mathematical representation of topics in text, as indicated by coherence metrics (Grootendorst 2022), many researchers posit that LDA models produce more interpretable and accurate topics (Abuzayed and Al-Khalifa 2021). This suggests the existence of both a validation gap and a standardization gap between traditional models such as LDA and neural models like BERTopic and Top2Vec (Hoyle et al. 2021). This disparity in evaluation methods is a common challenge in interdisciplinary research. Hence, in our initial exploration, we compared outputs from both BERTopic and guided LDA (Jagarlamudi, Daumé III, and Udupa 2012) and found that BERTopic yielded more interpretable results for our dataset, while guided LDA provided overly broad representations for our analysis.

We used guided BERTopic with sentence embeddings (all-MiniLM-L6-v2) (Reimers and Gurevych 2019) coupled with dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018). Using grid search, we assessed model sensitivity with three UMAP parameters: figure 2 demonstrates the highest *coherence score* with the *number of neighbors = 5*, *number of components = 5*, and *minimum distance = 0.6*. We also detected the optimal value for the *diversity parameter = 0.2* of Maximal Marginal Relevance (Carbonell and Goldstein 1998) used as a model that fine-tunes the topic representations of BERTopic.

In addition to BERTopic, we leveraged an alternative method known as Topic Detection Based on Semantics (Cheng et al. 2020) to automatically associate documents with a predefined ontology (Makkonen, Ahonen-Myka, and Salmenkivi 2004). In this approach, we utilized sentence transformers to generate text embeddings and matched each document in the dataset with one of the predefined topics by calculating the maximum cosine similarity. Unlike BERTopic, which generates topics with additional attention to seed words, the similarity-based method identifies topics that match the seed topics. All topic modeling experiments were done via the free version of Google Colab.

Finally, following previous literature (Capozzi et al. 2020), we employed the odds ratio metric to identify keywords characteristic to documents published after the introduction of the AIA proposal in April 2021 and the release of ChatGPT in late November 2022. This metric assesses the frequency of keywords in documents from the target period relative to their occurrence in other periods. To ensure reliability, we excluded keywords with low frequencies by setting a threshold of one standard deviation.

Computational approaches were complemented with qualitative content analysis to provide contextual insights. Several studies suggest that complementing qualitative anal-

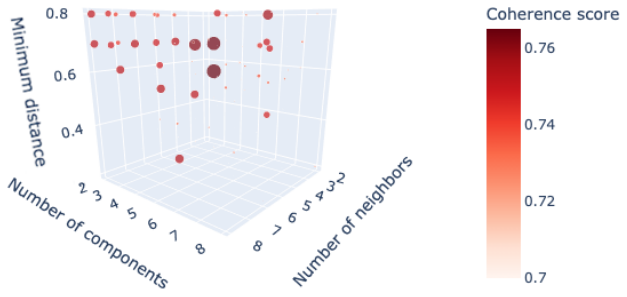


Figure 2: Sensitivity of UMAP parameters. Larger size and darker color of points indicate higher coherence.

ysis with computational methods provides an advantage of scaling while, given appropriate evaluation, maintaining integrity in the analysis process (Isoaho, Gritsenko, and Mäkelä 2021; Aranda et al. 2021). Qualitative coding data was selected in two ways: (1) passages containing compiled keywords were extracted from the corpus, and (2) 400 documents were proportionally sampled across the time period for data-driven coding. Both sets were analyzed for mentions of AI risks, with particular attention to GPAI. Risk-type codes were adapted from the taxonomies of Weidinger et al. (2022) and Bird, Ungless, and Kasirzadeh (2023).

The main coding was done by one of the authors, and a 10% sample of the coded documents was independently coded by two other researchers. Inter-coder reliability was assessed using an *averaged Cohen's Kappa metric = 0.81* and *percentual agreement = 87.7%* for binary agreement (whether or not any mention of a risk was present in the text). Multi-code agreement for different risk types was measured using the *Fleiss's Kappa = 0.73* and *Krippendorff's Alpha = 0.71* coefficients. The results of the content analysis were checked, negotiated, and synthesized into emerging key risk-related topics based on their salience in the examined sample. Altogether, passages from approximately 650 documents across all sources were included in the analysis.

Results and Analysis

Across all source types covering EU AI policy, we observe an expected increase in attention to GPAI following the release of ChatGPT in November 2022. Discussions of security and regulation around AI have grown in conjunction with its popularization. The following subsections present findings from the computational analysis using seed-guided topic models as well as qualitative analysis.

Computational Analysis

To compile seed words for topics related to AI risks, we referred to the taxonomies of risks posed by LLMs (Weidinger et al. 2022) and TTI models (Bird, Ungless, and Kasirzadeh 2023) as well as the AI HLEG guidelines (High-Level Expert Group on AI 2019). The topics "Discrimination, hate speech, and exclusion", "Mis- and disinformation", "Environment and sustainability", and "Cyber threats and malicious use" were directly derived from the taxonomies, with minor changes to topic names. Information hazards (Weidinger et al. 2022) was framed as "Privacy and personal information" to enhance topic coherence and distinction. Similarly, disinformation and misinformation were grouped as one topic as per Bird, Ungless, and Kasirzadeh (2023), and malicious uses (Weidinger et al. 2022; Bird, Ungless, and Kasirzadeh 2023) was targeted as a cybersecurity-related topic.

Other categories, such as socioeconomic harms (Weidinger et al. 2022), presented challenges due to broad and overlapping themes. We opted to target socioeconomic harms with seed words such as "job," "workforce," and "public sector" to capture AI's impact on employment and socioeconomic structures. Similarly, human-computer interaction harms (Weidinger et al. 2022), which includes user experience and psychological impacts, were not isolated due to their context-dependent nature. These complexities were considered during qualitative analysis and examination of the content of the produced topics to account for broader risk discourses while maintaining the topics actionable and coherent. The topics and keywords are provided in Table 2.

The seeded topic model achieved a coherence coefficient of 77% with 5-fold cross validation of 76%. A significance test based on topic significance in generative models (Al-Sumait et al. 2009) between the presented topics and random ones shows that the topics are significant (*p-value = 0.00002*). The topics are presented in Table 3, and their titles are based on representations. The modeling revealed topics associated with "Media", "Research Institute", "Robotics", and "Healthcare" as well as prominent topics related to "AI Systems" and "AI Regulation". It should be noted that some topics feature names of specific organizations (ELLIS for "Research Institute").

Over time, the publication of documents related to "AI Systems" and "AI Regulation" exhibits similar patterns in different scales, and "Media"-related topics show regular smaller peaks, most notably around January 2023. "Research Institute"-themed documents have small peaks in September 2021 and May 2023. In contrast, the remaining two topics show less popularity and lack strong representation throughout the analyzed period. The complete distribution is illustrated in the accompanying Figure 3.

To analyze the distribution of topic-related texts across sources on a timeline, we divided them into two categories: news articles and EU-related documents. The findings are presented in Figure 4. The predominant topic across both types of documents is "AI Systems". For the news article set, the second most popular topic is "Media", while "AI Regulation" and "Research Institute" are more prominent in the EU-related documents. In both sets, the number of documents related to the most popular topics increases during the latter half of 2021 and the early months of 2023. There is a peak in autumn 2021 among EU-related documents, specifically related to "AI Systems", "AI Regulation", "Research Institute", and "Robotics".

This peak appears to be explained by the publication ac-

topic	keywords / key phrases
Cyber threats and malicious use	attack, cyber, cybersecurity, malicious, adversarial
Discrimination, hate speech, and exclusion	bias, discrimination, non-discrimination, fair, fairness, exclusion, hate speech
Environment and sustainability	environmental, environment, sustainable, sustainability, climate, ecological
Mis- and disinformation	misinformation, disinformation, fake news
Labor market	labor, job, work, workforce, public service, public sector
Privacy and personal information	personal information, sensitive information, private, privacy, sensitive data, personal data

Table 2: Seed terms for guided topic modeling.

tivity around that time, which is at its highest in October 2021. The "Healthcare" topic captures documents related to public health services, discussions related to the COVID-19 crisis, the AIA requirement for the protection of health and safety, and AI innovation for healthcare. Although AI is largely presented as a driver for beneficial development in health services, risk assessment and management, as well as data protection in accordance with GDPR, are raised as private companies gain access to EU citizens' personal health data. "Robotics", on the other hand, is a regularly occurring theme in EU policy discussions both pre- and post-GPT. While it has crossover with GPAI, it is mostly mentioned as a growth area adjacent to rather than part of AI.

The timeline of news article topics reflects the rapid reactive publication cycle of news media, with a gradual climb in topic popularity in conjunction with increased discussion around the AIA and AI. The first peak in the topic "AI Regulation" occurs around the release of the first version of the AIA in April 2021. Another moderate increase in news attention is seen in May 2022, when the French Presidency of the CE proposed changes to the AIA text to regulate GPAI systems. After the release of ChatGPT in November 2022, the "AI Systems" and "AI Regulation" topics continue to grow, peaking around the time of the EP adoption of their negotiating position on the AIA in June 2023. The rise in "AI Regulation" in November 2022 appears correlated with the EU Council agreement on a final compromised version of the AIA, while the peak in February and March 2023 coincides with the EU Parliament's vote on the final version of the legal text. Active periods in the EU legislative trilogue process appear to be reflected in topic prominence.

Overall, while the generated topics reveal temporal trends in AI discourses, they show moderate alignment with the risk-related seed terms. This suggests that discourses specific to AI risks are dispersed across multiple predominant topics.

Next, to gain more targeted insight on our seed topics, we categorized all documents based on their semantic similarity to the seed terms. We determined the most related seed topic by calculating the maximum cosine similarity between the embeddings of the document and the embeddings of the seed terms. When using similarity-based detection of document topics, "Cyber threats and malicious use" emerges as the most popular topic, closely followed by "Privacy and per-

topic	representation
AI Systems	digital, new, systems, act, public, intelligence, artificial, said, use, artificial intelligence
AI Regulation	act, systems, intelligence, artificial, artificial intelligence, new, high risk, regulation, new, use
Media	media, digital, new, week, read, act, platforms, tech, said, countries
Research Institute	ellis, research, learning, university, machine learning, machine, unit, researchers, phd, program
Robotics	watch, robotics, public, report, standards, sector, robots, industrial, research, development
Healthcare	health, medical, healthcare, cancer, research, europe, patients, technologies, brain, covid19

Table 3: Results of guided topic modeling based on risk keywords applied to all collected documents.

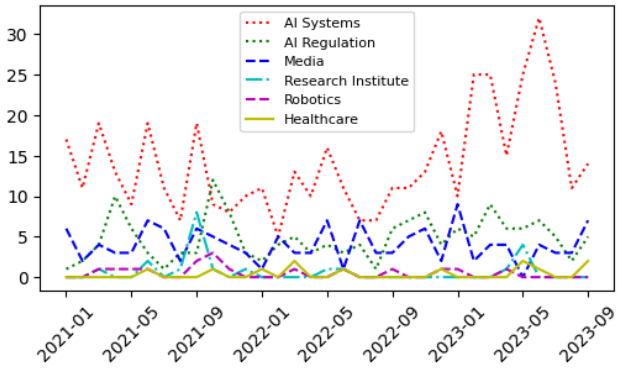
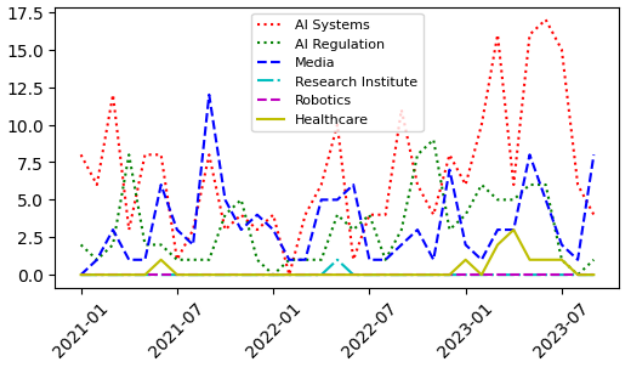
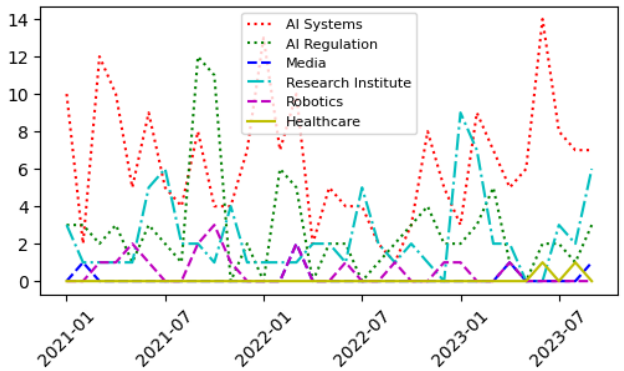


Figure 3: The monthly distribution of documents related to detected topics based on risk keywords.



(a) News media



(b) EU-related organizations

Figure 4: The monthly distribution of documents related to detected topics per source type.

sonal information". The distribution of documents related to both of these popular topics mirrors the patterns observed in "AI Systems"- and "AI Regulation"-related documents detected in topic modeling. The topic "Environment and sustainability" maintains a consistent presence throughout the analyzed period, along with "Labor market". The results are presented in Figure 5.

We divided the data into three periods: before the release of the AIA proposal, after its release, and after the launch of ChatGPT. Topics emerging in each period were explored using top keywords identified via odds ratio, which compares the likelihood of keyword occurrence in one period relative to others. During preprocessing, we excluded low-frequency as well as location- and time-related keywords. The top 15 keywords for each network are presented in Table 4.

The keywords reflect the discursive shifts following the two events and correlate with the topic modeling results. In the periods after the introduction of the AIA proposal, we observed an increase in regulation-related topics reflected in keywords such as "legislation", "gdpr", "standardisation", "lawmakers", "regulatory", and "compliance". The term "sanctions", which appears post-AIA, is mainly used in the context of non-compliance in the EU-related document set, while news organizations refer also to political or trade sanctions and "tariffs". A similar change is visible following the release of ChatGPT as the regulatory topics were joined by keywords such as "openai", "chatbot", and "generative". "Cybersecurity", "disinformation", and "robustness" emerged relative to the new challenges posed by rapid GPAI adoption.

Finally, we examined the texts from each period to identify defining patterns or characteristics. We utilized word shifts to estimate the differences among the word sets for each time slot within the collected documents, employing *shifterator* (Gallagher et al. 2021), a tool that generates word shift graphs illustrating how the words contribute to the disparities between two sets of texts, using a weighted measure. Specifically, we applied Jensen-Shannon divergence shifts due to their simplicity and ease of interpretation. Figure 6 shows the words shifts occurring after the introduction of the AIA proposal and after the release of ChatGPT. These shifts

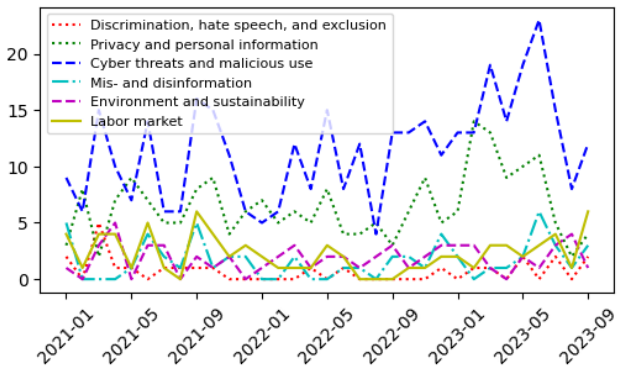


Figure 5: The monthly distribution of documents related to detected similarity-based topics in all data sources.

mirror trends seen in topic models and odds ratios, showing an early rise in regulatory discourse followed by increased attention to GPAI reflected in words such as "model", "generative", "chatgpt", "foundation", and "gpai". The keyword "swedish" relates to the Swedish Presidency of the CE in the first half of 2023, while Slovenia held that position in the second half of 2021 (reflected in the decline of the keyword "slovenian").

Qualitative Analysis

The qualitative content analysis aimed to complement the computational analyses by providing a more contextual insight into AI risk discourse. To study how the uptake of GPAI has affected EU AI policy discourse in terms of framing AI risks and security concerns, we synthesized four salient topics associated with risks: (a) cybersecurity; (b) education, upskilling, and AI literacy; (c) access to computational resources; and (d) the impact of scale.

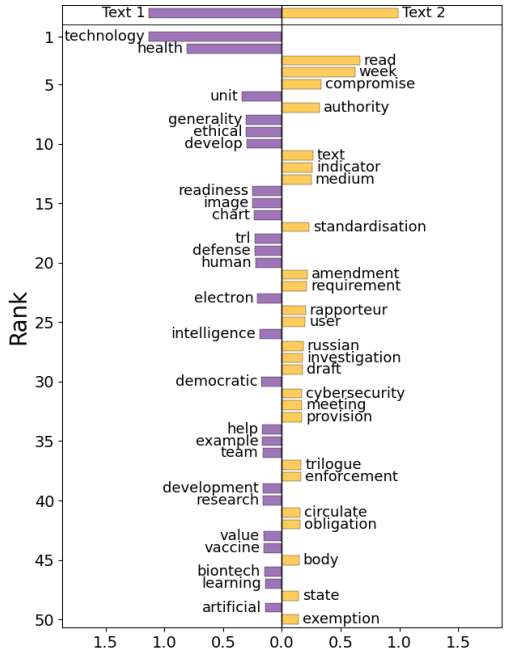
Cybersecurity and education emerged as prominent topics consistent with the results of the computational analysis. The other two, access to computational resources and the impact of scale, were identified as overarching topics in the discussion around GPAI. While media-related topics such as misinformation dissemination in news were raised as points of concern, they were typically characterized by a focus on AI and media literacy. The following subsections describe findings related to these risk topics. Environmental impacts of GPAI did not constitute a salient topic, which also corresponds to the topic model output.

Cybersecurity Discourse around cybersecurity highlights resilience against malicious AI use and the implementation of improved measures for robustness, privacy, and protection of sensitive or proprietary information. AI is seen to present new concerns undermining general and individual safety and security. In October 2023, the G7 countries agreed on the so-called Hiroshima Process including a set of 11 International Guiding Principles on Artificial Intelligence (The Group of Seven (G7) 2023), building on previous principles presented

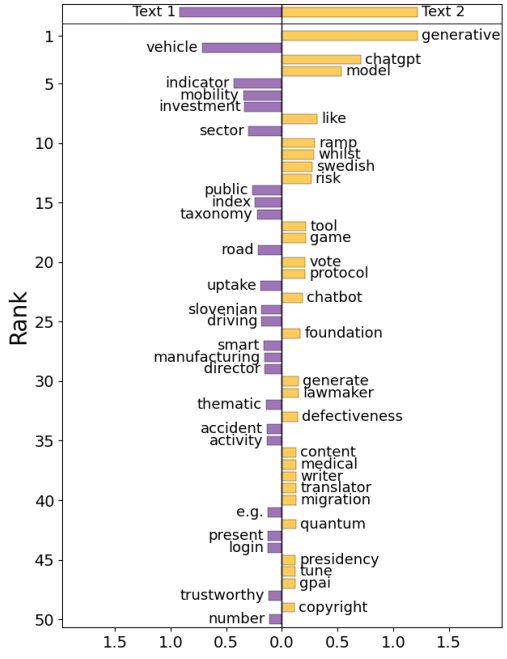
pre-AIA	post-AIA	post-ChatGPT
ethical	intelligence	legislation
digitalisation	international	gdpr
autonomous	automation	openai
robotics	agenda	chatbot
technologies	sanctions	standardisation
discriminatory	interoperability	generative
regulation	regulatory	chatgpt
policymakers	industrial	requirements
initiative	ethical	lawmakers
risks	ellis	brussels
biographer	europa	microsoft
harms	tariffs	compliance
readiness	intouchai	robustness
protection	procurement	disinformation
future	institute	cybersecurity

Table 4: Top 15 keywords by odds ratio for each period compared to other periods.

by the Organization for Economic Cooperation and Development (OECD). The principles include the proposition to "*invest in and implement robust security controls, including physical security, cybersecurity, and insider-threat safeguards across the AI lifecycle*" (AI Alliance).



(a) *before (left)* and *after (right) the AIA proposal release*



(b) *after the AIA proposal (left)* and *ChatGPT (right) releases*

Figure 6: Jensen-Shannon divergence shifts of words. The bars at the top show the overall difference.

The resilience of machine-learning algorithms and AI against adversarial learning, data manipulation, and data poisoning is a point of concern, as demonstrated in the report for Artificial Intelligence Cybersecurity Challenges by the European Union Agency for Cybersecurity (ENISA). ENISA produces an annual taxonomy of AI threats. For the year 2023, the threat landscape highlighted eight threats; ransomware, malware, social engineering, threats against data, threats against availability: denial of service, threats against availability: internet threats, information manipulation, and supply chain attacks (Lella et al. 2023). In a joint report from Stanford and Georgetown universities, they posit that *"the wider use of generative AI models like ChatGPT could make it hard to uncover information operations, as malicious actors could use them to produce highly convincing propaganda at a mass scale"* (AI Alliance). Suggestions were made to introduce new regulatory frameworks that mitigate threats across the entire AI lifecycle with collaborative efforts between technology industry stakeholders. A balance of innovation and risk mitigation is voiced by governments, EU organizations, and industry representatives alike. Civil society organizations such as European Digital Rights (EDRi) highlight concerns about the role of AI governance in regulating surveillance and safeguarding data privacy.

Cybersecurity is a growing area of interest due to its perceived benefits and, on the other hand, the need to sustain robust cyber infrastructures in relation to perceived threats. Especially in the context of cyber warfare, the benefits of AI for the protection of critical infrastructures, detection of cyber threats, and development of robust defensive techniques are widely recognized. However, while product safety and security remain crucial for AI providers, maintaining EU-internal control of AI lifecycles is raised as a security issue.

Education, Upskilling, and AI Literacy There is a growing emphasis on interdisciplinarity in research, education, and development teams in the industry, highlighting the importance of collaboration between diverse stakeholders. The notion of cross-sectoral collaboration reaches beyond AI design and development to participation in devising standardization measures. Despite these calls for cross-cutting collaboration between stakeholders, there is also a "*need for a network of research laboratories independent of industry interests, devoted to open AI research focused on finding solutions to these risks is increasingly apparent*" (ELLIS).

GPAI appears to partly shift the focus of risk toward end-user activity, positioning users not merely as recipients but as active participants in the AI lifecycle. As an active generator of content, the end-user is conceived either as a malicious actor or as a victim of malicious entities or black-boxed algorithms. When not framed as malicious actors, users are presented as navigating complex AI systems with the risk of being misled by dark patterns, anthropomorphic system design, or invisible data exchange between applications. Non-expert users *en masse* would then pose a range of risks related to information integrity: "*users do not care how these models generate language. They perceive it as human-like and they will often perceive it as authoritative*"

(EUobserver). The popular perception of GPAI tends to be "*disproportionately dominated by fears of AI-related existential threats*" (ELLIS) or concerns that "*powerful, 'intelligent' technologies will radically and unpredictably transform our reality — or even develop some form of life of their own*" (EUobserver). AI literacy thus becomes a part of risk mitigation.

There are calls for boosting digital, algorithmic, and media literacies, as well as support for upskilling of the general workforce, especially public servants and those working with critical infrastructures. Here, AI literacy becomes an essential building block, as education is seen to "*fuel the development*" of AI (AI4EU). Public sectors, which largely rely on privately developed AI systems, have increased responsibility in staying informed and up-to-date both to prevent AI-driven harm and to promote cost-effective digitalization through AI adoption and innovation. Furthermore, AI is presented to pose significant impacts on the labor market with demographically uneven consequences. Changes in the structure of the job market may place those more vulnerable at more risk, and the rapid adoption of GPAI has further fueled the discussion of who is at risk for job loss.

Access to Computational Resources We observed recurring discussions on access to compute in the EU. As models grow larger, more capable, and more interoperable, few actors have the possibility to develop state-of-the-art GPAI systems. Computational resources and hardware infrastructure emerge as key themes in EU policy discourse as enablers for the region to maintain a competitive position in AI. With the majority of resources concentrated outside the EU, there are initiatives calling for the EU to invest in technological self-reliance, promoting initiatives that support local compute, competitiveness, the ability to ensure trustworthy solutions, and cultural and linguistic diversity. Critiques are raised due to the EU's current reliance on imported technology and skills, which "*in a sector as crucial as AI, is a vulnerability that cannot be countenanced. Holding the world's most advanced regulatory framework for artificial intelligence bears scant worth if we are mere consumers of the technology we seek to regulate*" (AI Alliance). In a situation of unequally distributed compute, "*the window to act meaningfully from Europe often can appear vanishingly small*", while there are "*major future ethical and societal dilemmas linked to algorithmic power*" (EUobserver).

Impact of Scale The impact of a significant growth in scale is frequently mentioned in relation to most of the presented risks posed by GPAI. Increased accessibility and computational capability entail larger-scale content generation capacity. The ability to deploy systems at scale creates potential novel challenges for the integrity of democratic processes, the efficiency of crisis management, the maintenance of critical infrastructures, and civic access to essential services. For regulators, this marks a race "*to catch up with the speed of development of new technologies, such as ChatGPT*" (EUobserver). Governance is faced with "*the wildfire spread of GPAI*, which *"already marks a failure of European regulatory efforts*" (EUobserver). Scale, therefore, also exacerbates the challenges of future-proofing for AI regulation. Similarly to cybersecurity concerns, providers and policymakers both recognize these risks. Stakeholder collaboration and information sharing for policy development is needed to understand the inner workings of GPAI and to balance overly defensive measures leading to financial losses and regulatory costs.

GPAI is introducing changes to economic structures by lowering barriers to AI access, enabling small and medium-sized enterprises (SMEs) and individuals to use models for downstream tasks without substantial infrastructure costs. This may benefit entrepreneurial innovation and research, a priority for policymakers as "*language models are also an informal indicator of countries' advancement in AI research*" (Politico Europe). However, wide-scale accessibility is also seen to amplify risks of disinformation, fraud, or cyberattacks as policymakers brace for an "*explosion of an unregulated, unchecked technology able to churn out disinformation at scale and spout hateful, incorrect or biased content*" (Politico Europe). The growing scale of GPAI also raises questions about economic dependency and the EU's ability to compete in AI development while maintaining alignment with the values embedded in its regulation.

Discussion

The primary focus of this paper was to study how the proliferation of GPAI across society has affected the discourse surrounding EU AI policy in terms of AI risks and security. The discursive lens provides insights on how language and framing emerge as a sensemaking process for uncertain, novel, or disruptive technologies in society and reveals what kind of discourses are shaping current and future policy views (Ulnicane et al. 2021, 2022). Our chosen timeframe contextualized the analysis around the transition into the "post-GPT" era, where simultaneously the globally first AI-specific regulation was under negotiation in the EU co-legislative trialogue process. We generated topics of interest using topic modeling and other computational approaches in conjunction with qualitative content analysis and used the results to examine the discursive shift that has taken place post-GPT.

The computational analysis revealed a growing trend in regulation- and GPAI-related topics since the introduction of the AIA proposal and the release of ChatGPT, with cybersecurity, privacy, research, and education emerging as salient topics. In the qualitative analysis, we distilled central risk themes around the increased uptake of GPAI: cybersecurity; computational resources; education, upskilling, and AI literacy; and the scale of GPAI impacts. The risks outlined in the taxonomies of Weidinger et al. (2022) and Bird, Ungless, and Kasirzadeh (2023) are mentioned, but with GPAI, we observed an emphatic shift from the design and development phases onto the post-deployment phase. The risk mitigation discussion has therefore moved towards post-market monitoring as preemptive requirements for model transparency have begun to prove insufficient and models have become more and more accessible to the public. There is an increased focus on how those models and systems are used and by whom. The risks attributed to scale are particularly highlighted as the number of potential content-generating

stakeholders grows. Citizen-driven AI deployment necessitates a balance between supporting opportunities for innovation and maintaining safe and secure systems while ensuring inclusive access to AI and preventing market concentration. AI accessibility may lead to decentralized innovations that challenge current regulatory approaches including the AI Act, requiring the EU to adapt to an environment where AI is driven both by technology companies and individuals.

Although attention to cyber threats has obvious general gravity, cybersecurity also represents a rhetorically tangible way to justify security concerns posed by AI. Similarly, the concentration of most computing outside of the EU provides a justificatory angle highlighting economic and cyber risks such as those stemming from foreign data ownership and lack of transparency in model pipelines. To these ends, investing in AI education and boosting AI literacy are ways to build and maintain an informed public for information resilience. The public sector has a special role in facilitating upskilling given their responsibility for ensuring accessible services. Public AI services are expected to not compromise efficiency while remaining adaptable to technological change and implementing appropriate privacy measures.

While discriminative decision-making systems used for tasks such as assessment of creditworthiness or allocation of social benefits have been a source of concern for some time, growing generative capabilities have expanded the pool of risks. Although issues around the amplification of harmful and discriminatory social structures resulting from biased data and inadequate model training have neither disappeared nor become irrelevant, as exemplified by Weidinger et al. (2022) and Bird, Ungless, and Kasirzadeh (2023), GPAI is presented as more transformative and more *prototypically AI-like* in its outputs. With GPAI, the focal point of risk discourse is somewhat less on bias or over- and underfitting during model training, and more on human-machine interaction, wide-scale accessibility, transparency of generated content, and the pervasive impact of scale. Pre-trained open-source models allow downstream developers to skip or disregard the design and development stages altogether. Moreover, API-based access to GPAI models limits the possibility for downstream developers to fully engage in model and system design and development.

AI policy discourse is part of the collective process of making sense of what AI is, what it can do, what implications it has, and what should be done about it. The competitive dynamics involved in AI and its governance also highlighted by, *inter alia*, Smuha (2021); Ulnicane (2022); Bradford (2023) appear particularly highlighted in the GPAI lifecycle. Devising regulatory measures for AI entails framing a legitimate set of risks to various aspects of security. AI securitization occurs in cooperation with various organizations and the public, echoed in reactions to governance moves such as the AI Act (Mügge 2023), the EC AI strategy, or the AI HLEG guidelines. It should be noted that since the AI discourse is deeply rooted in industry narratives (Cath 2018; Nordström 2022), AI policy is a collaborative effort between policymakers and industry actors. The AI HLEG is a concrete example of this, with its significant number of industry members (Nordström 2022). The discourse that occurs around AI security therefore unfolds in multi-stakeholder discussions and remains in continuous flux (Eroukhmanoff 2017).

The findings are shaped by the regulatory and socio-political context of the EU. While discussions specific to the AIA are less generalizable outside the EU, they may have international implications through the 'Brussels effect' (Bradford 2020), a phenomenon where EU regulatory policies and standards influence global practices. The evolving role of users, the increased accessibility of models, the need for AI literacy, and cybersecurity concerns pose challenges worldwide regardless of regulatory compliance. These topics could therefore be further studied in, for example, US and Chinese contexts and companies.

Regulations tend to play catch-up with developments in AI due to the nature and out-of-sync timelines of lawmaking versus profitable technological progress. The results show that there are layers in the narrative that seem to steer policy, and this narrative fluctuates with technological developments and multi-stakeholder opinions as expressed in the sources examined in this study.

Limitations

Like in all research, there are limitations that should be considered when examining the results. First, the data may contain some inherent biases. Media sources may prioritize risk narratives that resonate with their audiences, while policy reports reflect institutional priorities. The forum and blog posts, while offering more grassroots perspectives, are likely skewed toward politically and technologically literate demographics, potentially underrepresenting non-English speakers, marginalized groups, or smaller EU member states.

Secondly, considering the breadth of the data, AI securitization is discussed here as a discursive process between various stakeholders, even if instigated by EU policymakers. The results are therefore not indicative of an official EU AI policy agenda or deliberate rhetorical strategy but instead offer a perspective to wider policy-related discourse. Methodologically, both qualitative reading and topic models have limitations, the former relying on researcher interpretation and the latter on statistical word co-occurrence with limited context. The quality of topics produced relies on assumptions and parameters set by the researcher, which can lead to biases, and topic models may struggle with nuanced semantics, potentially oversimplifying the findings. We have attempted to address these issues by combining computational methods with qualitative analysis.

Conclusion

We examined the presentation of risks posed by the development and adoption of GPAI in EU policy-related texts and EU policy journalism using guided topic modeling and qualitative analysis. The data, situated around the transition to a "post-GPT" era, allowed us to identify shifts in AI risk discourse in intersection with developing governance.

Salient risk-related topics emerging around increased GPAI uptake included cybersecurity, regional imbalance in compute, education and AI literacy, and the overarching effect of these risks occurring at scale. Simultaneously, GPAI

discourse has introduced a shift in emphasis from the design and development of models and systems onto post-deployment and post-market monitoring.

The discussions around cybersecurity and EU preparedness for rapid and large-scale technological change emerge as justifications for a need for AI governance. While regulation remains presented as reactive in the evolving AI space, investing in AI research, education, and promotion of AI literacy are offered as tools for economic and technological security and information resilience.

Reproducibility: The data and code described in this paper are available for research purposes in the public repository https://github.com/vitiugin/eu-ai-discourse.

Acknowledgements

This research is part of the CAAI project funded by the Kone Foundation (202107233) and the Research Council of Finland (357349). A part of this work was supported by funding from JSPS KAKENHI (24K21058) and the SRC Trust-M project (353529).

References

Abuzayed, A.; and Al-Khalifa, H. 2021. BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Rechnique. *Procedia Computer Science*, 189: 191–194.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

af Malmborg, F. 2023. Narrative Dynamics in European Commission AI Policy—Sensemaking, Agency Construction, and Anchoring. *Review of Policy Research*, 40(5): 757–780.

Albalawi, R.; Yeap, T. H.; and Benyoucef, M. 2020. Using Topic Modeling Methods for Short-text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, 3: 42.

AlSumait, L.; Barbará, D.; Gentle, J.; and Domeniconi, C. 2009. Topic significance ranking of LDA generative models. In *ECML PKDD 2009, Bled, Slovenia, Proceedings, Part I 20*, 67–82. Springer.

Aranda, A. M.; Sele, K.; Etchanchu, H.; Guyt, J. Y.; and Vaara, E. 2021. From Big Data to Rich Theory: Integrating Critical Discourse Analysis with Structural Topic Modeling. *European Management Review*, 18(3): 197–214.

Axelborn, H.; and Berggren, J. 2023. Topic Modeling for Customer Insights: A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls. Master's thesis. Umeå University, Faculty of Science and Technology, Department of Mathematics and Mathematical Statistics.

Balzacq, T. 2011. *Securitization Theory: How Security Problems Emerge and Dissolve*. Routledge, Taylor & Francis.

Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of risks of generative text-to-image models. In *Proc. of the 2023 AAAI/ACM AIES*, 396–410.

Bradford, A. 2020. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press, USA.

Bradford, A. 2023. *Digital Empires: The Global Battle to Regulate Technology*. Oxford University Press.

Capozzi, A.; Francisci Morales, G. D.; Mejova, Y.; Monti, C.; Panisson, A.; and Paolotti, D. 2020. Facebook Ads: Politics of Migration in Italy. In *SocInfo*, 43–57. Springer.

Carbonell, J.; and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the ACM SIGIR*, 335–336.

Cath, C. 2018. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133): 20180080.

Cave, S.; Craig, C.; Dihal, K.; Dillon, S.; Montgomery, J.; Singler, B.; and Taylor, L. 2018. *Portrayals and Perceptions of AI and Why They Matter*. The Royal Society.

Cheng, P.; Du, J.; Kou, F.; Xue, Z.; and Chen, P. 2020. Topic Detection Based on Semantics, Time and Social Relationship. In *Proc. of CIAC*, 691–698. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Egger, R.; and Yu, J. 2021. Identifying Hidden Semantic Structures in Instagram Data: A Topic Modelling Comparison. *Tourism Review*, 77(4): 1234–1246.

Elish, M. C.; and Boyd, D. 2018. Situating Methods in the Magic of Big Data and AI. *Communication Monographs*, 85(1): 57–80.

Engler, A. C.; and Renda, A. 2022. *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act*. CEPS.

Eroukhmanoff, C. 2017. Securitisation Theory: An Introduction. In McGlinchey, S.; Walters, R.; and Sheinpflug, C., eds., *International Relations Theory*. Oxford: E-International Relations Publishing.

European Commission. 2021. 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. 21 April, Brussels.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. 2021. An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Ethics, Governance, and Policies in Artificial Intelligence*, 19–39.

Gallagher, R. J.; Frank, M. R.; Mitchell, L.; Schwartz, A. J.; Reagan, A. J.; Danforth, C. M.; and Dodds, P. S. 2021. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1): 4.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92.

Golpayegani, D.; Pandit, H. J.; and Lewis, D. 2022. AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. In *Towards a Knowledge-Aware AI: SEMANTiCS 2022, Vienna, Austria*, volume 55, 51. IOS Press.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI. European Commission, Brussels.

Hoffman, M.; Narayanan, M.; Mitra, A.; Liao, Y.-J.; and Frase, H. 2023. CSET AI Harm Taxonomy for AIID and Annotation Guide. 25 July, Center for Security and Emerging Technology (CSET).

Hoyle, A.; Goel, P.; Hian-Cheong, A.; Peskov, D.; Boyd-Graber, J.; and Resnik, P. 2021. Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. *NeurIPS*, 34: 2018–2033.

Isoaho, K.; Gritsenko, D.; and Mäkelä, E. 2021. Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Studies Journal*, 49(1): 300–324.

Jagarlamudi, J.; Daumé III, H.; and Udupa, R. 2012. Incorporating lexical priors into topic models. In *Proc. of the 13th Conference of the European Chapter of ACL*, 204–213.

Johnson, D. G.; and Verdicchio, M. 2017. Reframing AI Discourse. *Minds and Machines*, 27(4): 575–590.

Kassim, H.; Peterson, J.; Bauer, M. W.; Connolly, S.; Dehousse, R.; Hooghe, L.; and Thompson, A. 2013. *The European Commission of the Twenty-first Century*. OUP Oxford.

Laux, J. 2023. Institutionalised Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act. *AI & Society*.

Lella, I.; Ciobanu, C.; Tsekmezoglou, E.; Theocharidou, M.; Magonara, E.; Malatras, A.; Svetozarov Naydenov, R.; et al. 2023. ENISA threat landscape 2023: July 2022 to June 2023.

Li, C.; Chen, S.; Xing, J.; Sun, A.; and Ma, Z. 2018. Seed-guided topic model for document filtering and classification. *ACM TOIS*, 37(1): 1–37.

Makkonen, J.; Ahonen-Myka, H.; and Salmenkivi, M. 2004. Simple semantics in topic detection and tracking. *Information retrieval*, 7: 347–368.

McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In *Proc. of the AAAI Conference*, volume 35, 15458–15463.

McInnes, L.; Healy, J.; and Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

Mügge, D. 2023. The Securitization of the EU's Digital Tech Regulation. *Journal of European Public Policy*, 30(7): 1431–1446.

Nordström, M. 2022. AI under great uncertainty: implications and decision strategies for public policy. *AI & Society*, 37(4): 1703–1714.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of the ACL EMNLP*.

Slattery, P.; Saeri, A. K.; Grundy, E. A. C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2024. A systematic evidence review and common frame of reference for the risks from artificial intelligence. *ResearchGate Repository*. Preprint.

Smuha, N. A. 2021. From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence. *Law, Innovation and Technology*, 13(1): 57–84.

The Group of Seven (G7). 2023. Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI system.

Ulnicane, I. 2022. Artificial Intelligence in the European Union: Policy, Ethics and Regulation. In *The Routledge Handbook of European Integrations*. Taylor & Francis.

Ulnicane, I.; Knight, W.; Leach, T.; Stahl, B. C.; and Wanjiku, W.-G. 2021. Framing Governance for a Contested Emerging Technology: Insights from AI Policy. *Policy and Society*, 40(2): 158–177.

Ulnicane, I.; Knight, W.; Leach, T.; Stahl, B. C.; and Wanjiku, W.-G. 2022. Governance of Artificial Intelligence: Emerging International Trends and Policy Frames. In *The Global Politics of Artificial Intelligence*. Taylor & Francis.

Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach. *Computer Law Review International*, 22(4): 97–112.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proc. of the ACM FAccT*, 214–229.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Yurrita, M.; Murray-Rust, D.; Balayn, A.; and Bozzon, A. 2022. Towards a Multi-Stakeholder Value-Based Assessment Framework for Algorithmic Systems. In *Proc. of the ACM FAccT*, 535–563. New York, NY, USA.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this study provides new information without compromising ethics.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract summarizes the key contributions of the paper.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Methods.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Data.

(e) Did you describe the limitations of your work? Yes, see the Limitations.

(f) Did you discuss any potential negative societal impacts of your work? No, because the paper studies public discourse around AI.

(g) Did you discuss any potential misuse of your work? NA

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No, because the paper studies public discourse around AI.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, this paper conforms to the ethics review guidelines.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, check the Reproducibility.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see the Methods and code implementation.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, see the Methods.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the Methods.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Results and Analysis.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because only publicly available documents were used in the analysis.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see Wilkinson et al. (2016))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA