SHIELDING EU SECURITY: MITIGATING AI RISKS FOR STRATEGIC PROTECTION

Group 25
Annabel Duport, Jian Feng, Paul Arnold, Pirun Chan, Veronika Zadernivska

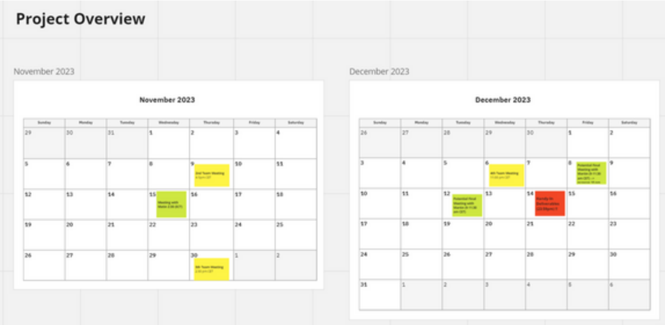
METHODOLOGICAL APPENDIX

For this project, we were assigned to work together on the topic of digitalization. The policy challenge regarding digitalization in 2023 states that, “AI has become increasingly pervasive across various sectors, leading to a remarkable transformation of industries and society. Despite its great potential, it presents risks and challenges such as ethical concerns and spread of mis(dis)information.” It raises the question, “What measures and policies should the EU put in place in order to ensure transparency, accountability, and privacy protection in AI systems?”

In order to effectively manage our project's development, we opted to utilize a Miro Board, which enabled us to track the progress of the project and monitor various tasks. Additionally, we utilised Google Docs to collaborate on the policy paper and review each other's work. Furthermore, we leveraged the convenience of a private Whatsapp group chat to facilitate seamless communication and task distribution, as well as coordinate group meetings based on our individual availabilities.

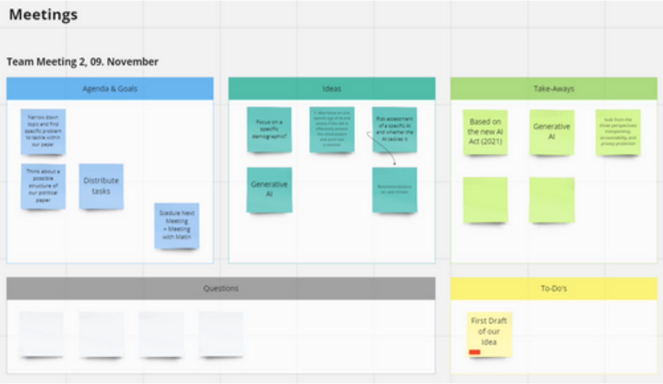
PROJECT MANAGEMENT

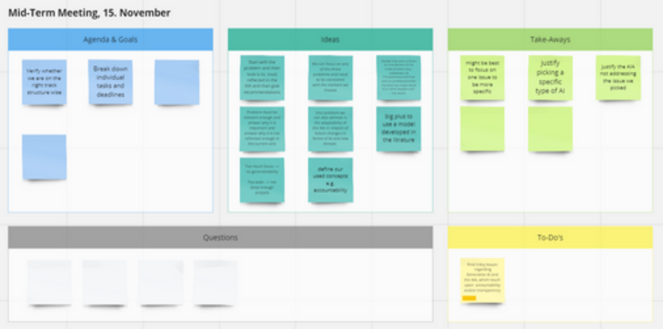
We began by establishing the crucial dates for this project, such as scheduled meetings with our supervisor Matin Mohaghegh, and the deadlines for submitting our deliverables. We also utilised this area to record the URLs of our collaborative documents to prevent any loss of our collective efforts.



MEETINGS

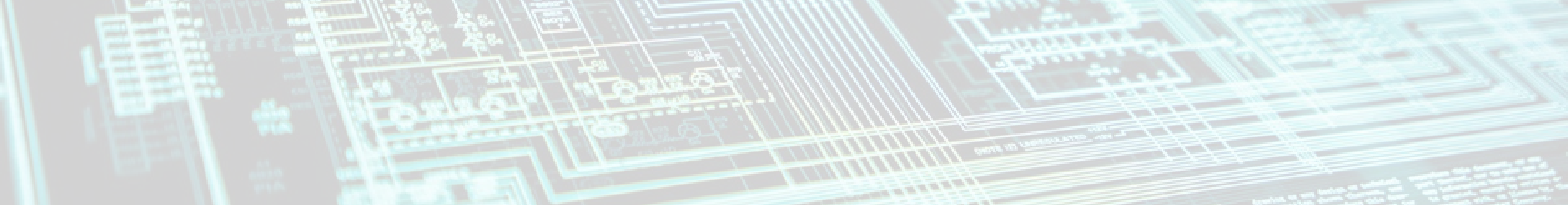
We organized our scheduled meetings in Waller; as a reserved meetings channel, we also used Zoom. Throughout our meetings, we utilized a helpful template, as displayed below, to establish objectives, topics for discussion and decision-making, noteworthy concepts arising from our dialogue, significant takeaways, and tasks to be completed prior to reconvening. In total, we held five meetings, including sessions with our supervisor. Supervisor`s meetings were always ending with setting the dates and tasks. After concluding our discussions, we would each work independently and reconvene later to report on our progress. During team meetings, we worked on presenting intermediate results, and research findings as well as distributing tasks and setting the agenda for the next meeting.

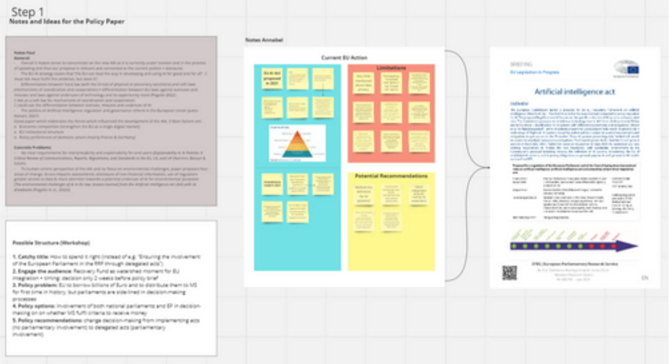

Meetings
Team Meeting 2, 09. November
Agenda & Goals
Ideas
Take-Aways
Narrow down topic and find specific problem to tackle within our paper
Think about a possible structure of our political paper
Distribute tasks
Scedule Next Meeting + Meeting with Matin
Focus on a specific demographic?
Generative AI
Risk assessment of a specific AI and whether the AI tackles it
Based on the new AI Act (2021)
Generative AI
Questions
To-Do's
First Draft of our Idea


Mid-Term Meeting, 15. November
Agenda & Goals
Ideas
Take-Aways
Verify whether we are on the right track structure wise
Break down individual tasks and deadlines
big plus to use a model developed in the literature
define our used concepts e.g. accountability
might be best to focus on one issue to be more specific
justify picking a specific type of AI
justify the AIA not addressing the issue we picked
Questions
To-Do's

STEP 1 - RESEARCH ON THE TOPIC

Following our initial meeting, we collectively decided to conduct individual research on digitalization and AI regulation. We aimed to become well-versed in the topic and develop informed perspectives and ideas on AI, its risks, and the challenges associated with the EU's current course of action. Each of us delved into the AI Act, which was deliberated by the EU Commission in 2021 and remains a subject of debate today. Through our research, we identified possible shortcomings and areas for improvement, which we intend to present as recommendations. This research was based on news articles mentioning the act, critiques, and evaluations of the paper by scholars and, of course, familiarising ourselves with the Act itself.



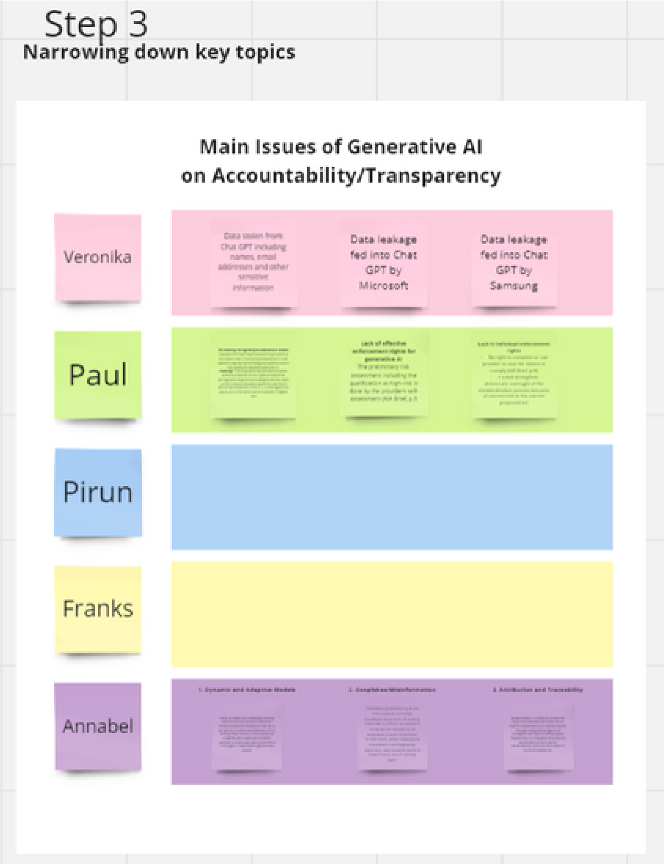

Step 1
Notes and Ideas for the Policy Paper
Notes Annabel
Current EU Action
Limitations
Potential Recommendations
Artificial intelligence act

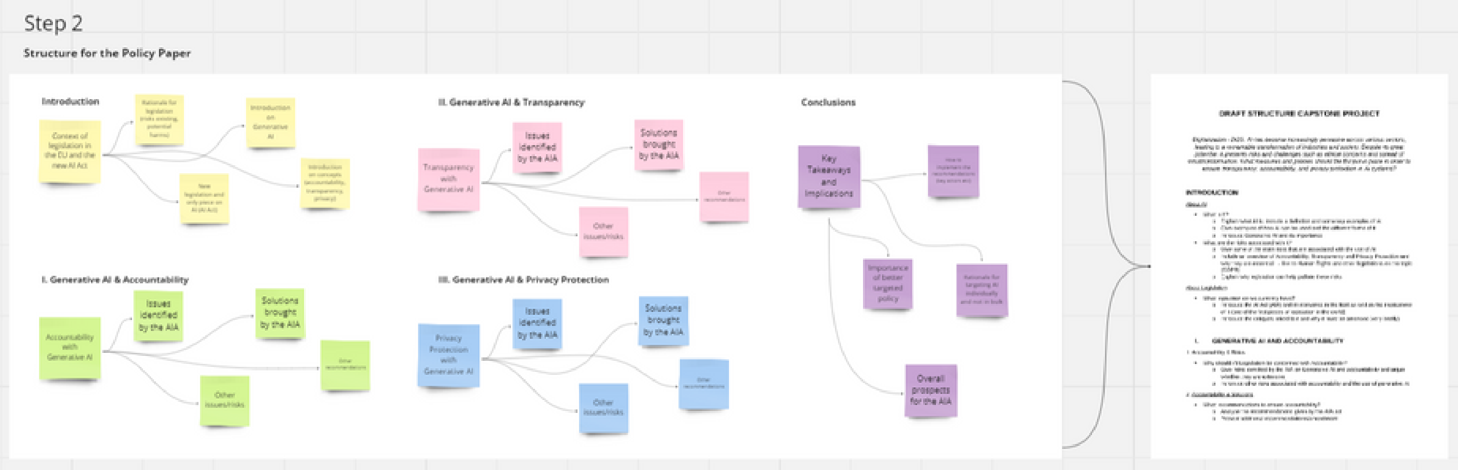
STEP 2 - STRUCTURING OUR POLICY PAPER

After achieving a thorough understanding of the topic, we proceeded to deliberate on the potential structure of our policy paper. The first deliverable was a rough common document, which later was reformatted into a first draft for supervision. The initial draft, which we have included below, served as the basis for our work. Although the structure underwent significant changes, it allowed us to concentrate on the important issues of accountability and transparency that we discovered during our research and discussions. We have always been in contact with our supervisor, Matin, who provided valuable and useful feedback on enhancing our paper further. Despite our desire to explore a broad range of topics, such as privacy protection, we soon realized it was not practical and we were not able to cover such a wide range of issues. This realisation prompted us to reassess our focus area. Thus, we were able to narrow down the main research area and proceed with our work.

STEP 3 - FOCUSING ON KEY AREAS

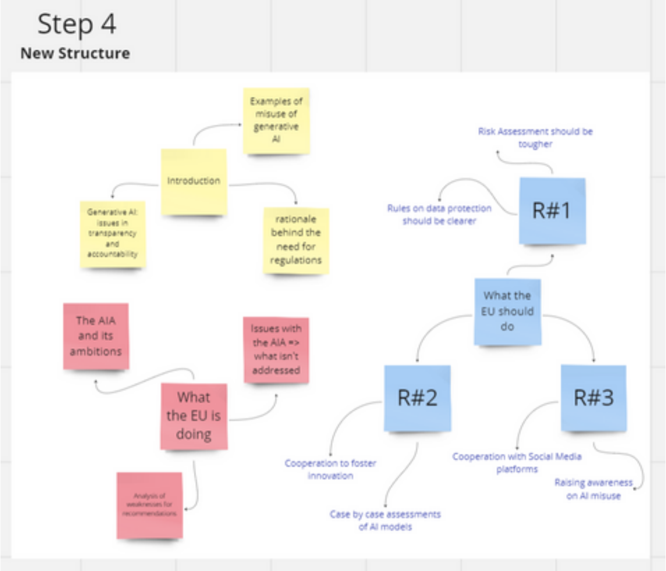
As previously discussed, we recognized the need to narrow our focus to provide more precise analysis and recommendations. To achieve this, we identified three critical areas of concern by presenting case studies that illustrated the potential pitfalls of AI technology, particularly regarding transparency and accountability. These examples served as a foundation for our revised approach.


Step 3
Narrowing down key topics
Main Issues of Generative AI on Accountability/Transparency
Veronika
Paul
Pirun
Franks
Annabel
Data stolen from Chat GPT including names, email addresses and other sensitive information
Data leakage fed into Chat GPT by Microsoft
Data leakage fed into Chat GPT by Samsung
Lack of effective enforcement rights for generative AI
Lack in individual enforcement rights
1. Dynamic and Adaptive Models
2. Deepfakes/Misinformation
3. Attribution and Traceability


Step 2
Structure for the Policy Paper
Introduction
Context of legislation in the EU and the new AI Act
Rationale for legislation (risks existing, potential harms)
Introduction on Generative AI
New legislation and only piece on AI (AI Act)
Introduction on concepts (accountability, transparency, privacy)
I. Generative AI & Accountability
Accountability with Generative AI
Issued identified by the AIA
Solutions brought by the AIA
Other recommendations
Other issues/risks
II. Generative AI & Transparency
Transparency with Generative AI
Issued identified by the AIA
Solutions brought by the AIA
Other recommendations
Other issues/risks
III. Generative AI & Privacy Protection
Privacy Protection with Generative AI
Issued identified by the AIA
Solutions brought by the AIA
Other recommendations
Other issues/risks
Conclusions
Key Takeaways and Implications
How to implement the recommendations
Importance of better targeted policy
Rationale for targeting AI individually and not in bulk
Overall prospects for the AIA
DRAFT STRUCTURE CAPSTONE PROJECT
INTRODUCTION
I. GENERATIVE AI AND ACCOUNTABILITY

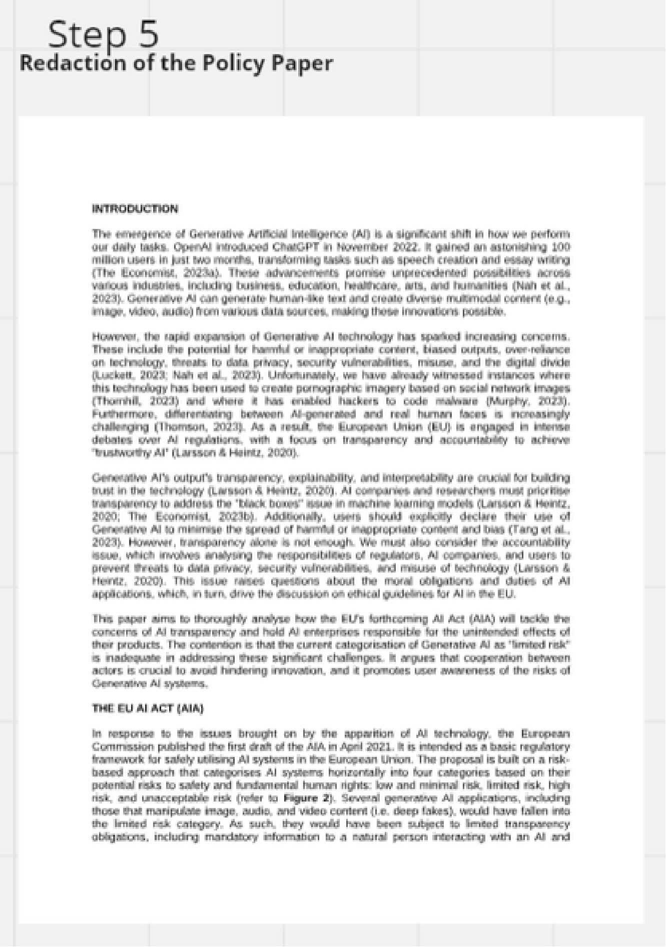
STEP 4 - RETHINKING THE STRUCTURE

Halfway into the project, we held meetings that gave us a clearer picture of our research problem and case studies that would help us better communicate the importance and relevance of our chosen topic. After carefully selecting the key examples and issues we wished to delve into, we realized that restructuring our paper would be necessary to hone in on our areas of interest. This revised format ultimately served as the foundation for the remainder of our research. It comprises three integral components: an introduction to the issue supplemented by illustrative examples, an analysis of the EU's response in the form of the AI Act, and finally, a set of recommendations for improving the current EU project.


Step 4
New Structure
Examples of misuse of generative AI
Introduction
Generative AI: issues in transparency and accountability
rationale behind the need for regulations
The AIA and its ambitions
Issues with the AIA => what isn't addressed
What the EU is doing
Analysis of weaknesses for recommendations
Risk Assessment should be tougher
Rules on data protection should be clearer
R#1
What the EU should do
R#2
R#3
Cooperation to foster innovation
Case by case assessments of AI models
Cooperation with Social Media platforms
Raising awareness on AI misuse

STEP 5 - WRITING THE POLICY PAPER

An important step for further work was to determine the final structure of the paper. Such discussions were conducted during online meetings, as well as with the help of online voting in the working chat. After agreeing on the structure, we proceeded to assign the various sections of the policy paper during a meeting, ensuring that the text would have a cohesive flow by blending our writing styles. With the first draft completed, we presented it to our supervisor, who provided the necessary feedback to help us refine and strengthen the paper. We incorporated this feedback into our work on the final details, striving to create a persuasive and unified policy paper.



Step 5

Redaction of the Policy Paper

INTRODUCTION

The emergence of Generative Artificial Intelligence (AI) is a significant shift in how we perform our daily tasks. OpenAI introduced ChatGPT in November 2022. It gained an astonishing 100 million users in just two months, transforming tasks such as speech creation and essay writing (The Economist, 2023a). These advancements promise unprecedented possibilities across various industries, including business, education, healthcare, arts, and humanities (Nah et al., 2023). Generative AI can generate human-like text and create diverse multimodal content (e.g., image, video, audio) from various data sources, making these innovations possible.

However, the rapid expansion of Generative AI technology has sparked increasing concerns. These include the potential for harmful or inappropriate content, biased outputs, over-reliance on technology, threats to data privacy, security vulnerabilities, misuse, and the digital divide (Luckett, 2023; Nah et al., 2023). Unfortunately, we have already witnessed instances where this technology has been used to create pornographic imagery based on social network images (Thornhill, 2023) and where it has enabled hackers to code malware (Murphy, 2023). Furthermore, differentiating between AI-generated and real human faces is increasingly challenging (Thomson, 2023). As a result, the European Union (EU) is engaged in intense debates over AI regulations, with a focus on transparency and accountability to achieve "trustworthy AI" (Larsson & Heintz, 2020).

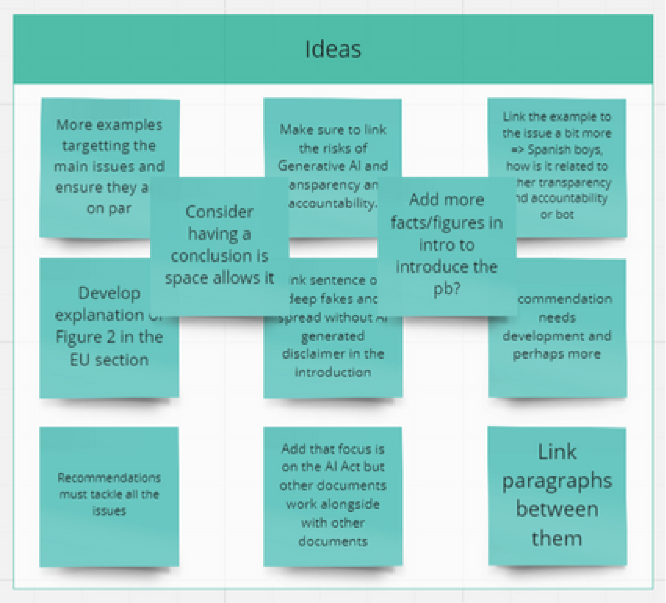
Generative AI's output's transparency, explainability, and interpretability are crucial for building trust in the technology (Larsson & Heintz, 2020). AI companies and researchers must prioritise transparency to address the "black boxes" issue in machine learning models (Larsson & Heintz, 2020; The Economist, 2023b). Additionally, users should explicitly declare their use of Generative AI to minimise the spread of harmful or inappropriate content and bias (Tang et al., 2023). However, transparency alone is not enough. We must also consider the accountability issue, which involves analysing the responsibilities of regulators, AI companies, and users to prevent threats to data privacy, security vulnerabilities, and misuse of technology (Larsson & Heintz, 2020). This issue raises questions about the moral obligations and duties of AI applications, which, in turn, drive the discussion on ethical guidelines for AI in the EU.

This paper aims to thoroughly analyse how the EU's forthcoming AI Act (AIA) will tackle the concerns of AI transparency and hold AI enterprises responsible for the unintended effects of their products. The contention is that the current categorisation of Generative AI as "limited risk" is inadequate in addressing these significant challenges. It argues that cooperation between actors is crucial to avoid hindering innovation, and it promotes user awareness of the risks of Generative AI systems.

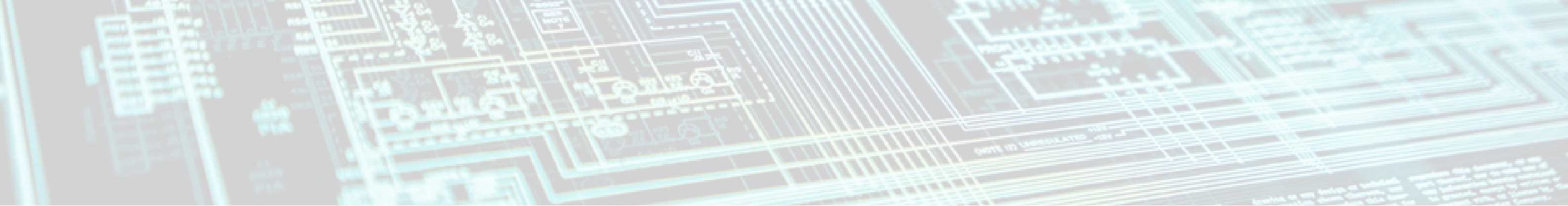
THE EU AI ACT (AIA)

In response to the issues brought on by the apparition of AI technology, the European Commission published the first draft of the AIA in April 2021. It is intended as a basic regulatory framework for safely utilising AI systems in the European Union. The proposal is built on a risk-based approach that categorises AI systems horizontally into four categories based on their potential risks to safety and fundamental human rights: low and minimal risk, limited risk, high risk, and unacceptable risk (refer to **Figure 2**). Several generative AI applications, including those that manipulate image, audio, and video content (i.e. deep fakes), would have fallen into the limited risk category. As such, they would have been subject to limited transparency obligations, including mandatory information to a natural person interacting with an AI and

After meetings and reflection on the received feedback, we again received a set of new ideas for our final paper.


Ideas
More examples targetting the main issues and ensure they a on par
Make sure to link the risks of Generative AI and ansparency an accountability.
Link the example to the issue a bit more => Spanish boys, how is it related to her transparency nd accountability or bot
Consider having a conclusion is space allows it
Add more facts/figures in intro to introduce the pb?
Develop explanation of Figure 2 in the EU section
nk sentence o deep fakes anc spread without AI generated disclaimer in the introduction
commendation needs development and perhaps more
Recommendations must tackle all the issues
Add that focus is on the AI Act but other documents work alongside with other documents
Link paragraphs between them

As a result of our cooperation, we received a well-developed, structured, and visually attractive policy paper, policy brief, and this methodological appendix.



MATERIALS USED DURING THE PROJECT

Here below, you can view a list of sources and literature that have helped us in writing this research.

Barani, M. and Van Dyke, P. (2023). *Generative AI and the EU AI Act - A Closer Look*. [online] Allen & Overy. Available at: https://www.allenovery.com/en-gb/global/blogs/tech-talk/generative-ai-and-the-eu-ai-act-a-closer-look.

European Commission (2022). *Regulatory framework on AI | Shaping Europe's digital future*. [online] digital-strategy.ec.europa.eu. Available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

European Parliament (2023). *Artificial Intelligence Act – Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. [online] European Parliament. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

Goldstein, A. (2019). *Teal LED Panel. Unsplash.com*. Available at: https://unsplash.com/photos/teal-led-panel-EUsVwEOsblE [Accessed 11 Dec. 2023].

Howarth, J. (2021). 80+ AI Stats: *Market Size, Growth & Business Use*. [online] Exploding Topics. Available at: https://explodingtopics.com/blog/ai-statistics.

Larsson, S. and Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review: Journal on Internet Regulation*. 9(2), pp.1-16.

Lawton, G. (2023). *How to prevent deepfakes in the era of generative AI | TechTarget*. [online] Security. Available at: https://www.techtarget.com/searchsecurity/tip/How-to-prevent-deepfakes-in-the-era-of-generative-AI.

Luckett. J. (2023). Regulating generative AI: a pathway to ethical and responsible implementation. *International Journal on Cybernetics and Informatics*. 12(5), pp.79-92.

Muhammed , S.T. and Mathew, S.K. (2022). The disaster of misinformation: a review of research in social media. *International Journal of Data Science and Analytics*, [online] 13(4). doi:https://doi.org/10.1007/s41060-022-00311-6.

Murphy, H. (2023). AI: a new tool for cyber attackers or defenders?. *Financial Times*. [Online]. 21 September. [Accessed 9 December 2023]. Available from: https://www.ft.com/content/09d163be-0a6e-48f8-8185-6e1ba1273f42.

Nah, F.F, Zheng, R., Cai, J., Siau, K. and Chen, L. (2023). Generative ai and chatgpt: applications, challenges, and ai-human collaboration. *Journal of Information Technology Case and Application Research*. 25(3), pp.277-304.

Stark, B., Magin, M., Geiß, S. (2021). *Meinungsbildung in und mit sozialen Medien. In: Schmidt, JH., Taddicken, M. (eds) Handbuch Soziale Medien. Springer Reference Sozialwissenschaften*. [online] Springer VS, Wiesbaden. doi:https://doi.org/10.1007/978-3-658-03895-3_23-1

Tang, A., Li, K., Kwok, K.O., Cao, L., Luong, S. and Tam, W. (2023). The importance of transparency: declaring the use of generative artificial intelligence (AI) in academic writing. *Journal of Nursing Scholarship*. 00, pp.1-5.

The Economist. (2023a). How to worry wisely about AI. *The Economist*. [Online]. 22 April. [Accessed 2 December 2023]. Available from: https://www.economist.com/leaders/2023/04/20/how-to-worry-wisely-about-artificial-intelligence.

The Economist. (2023b). How generative models could go wrong? *The Economist*.

Thornhill, J. 2023. The promise – and peril – of generative ai. *Financial Times*. [Online]. 28 September. [Accessed 2 December 2023]. Available from: https://www.ft.com/content/e6a391c7-bfd2-4eb1-82e5-6bc4eac9b131.

Ver Meer, D. (2023). *ChatGPT Stats: Key User Numbers, Revenue & Data*. [online] NamePepper. Available at: https://www.namepepper.com/chatgpt-users [Accessed 11 Dec. 2023].

Vittoriosi, E. (2023). Chat GPT. *Unsplash.com*. Available at: https://unsplash.com/photos/a-laptop-computer-sitting-on-top-of-a-wooden-table-G_vWviqUCCg [Accessed 11 Dec. 2023].