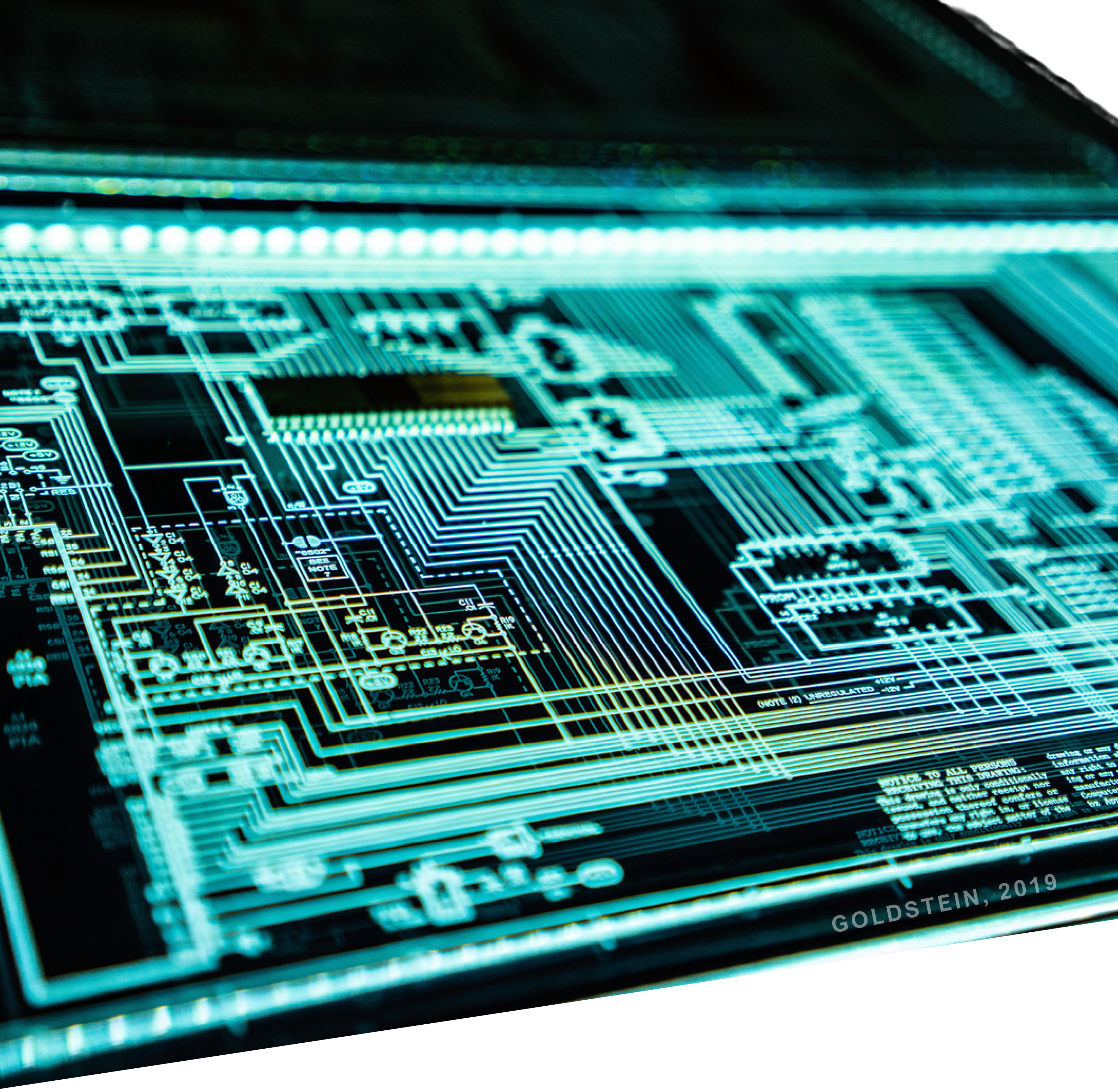
Annabel Duport
Jian Feng
Paul Arnold
Pirun Chan
Veronika Zadernivska

GROUP 25

ASSESSMENT COVER PAGE

DIGITALISATION DI23

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 the spread of mis(dis)information. What measures and policies should the EU put in place in order to ensure transparency, accountability, and privacy protection in the AI system?



GOLDSTEIN, 2019

POLICY PAPER

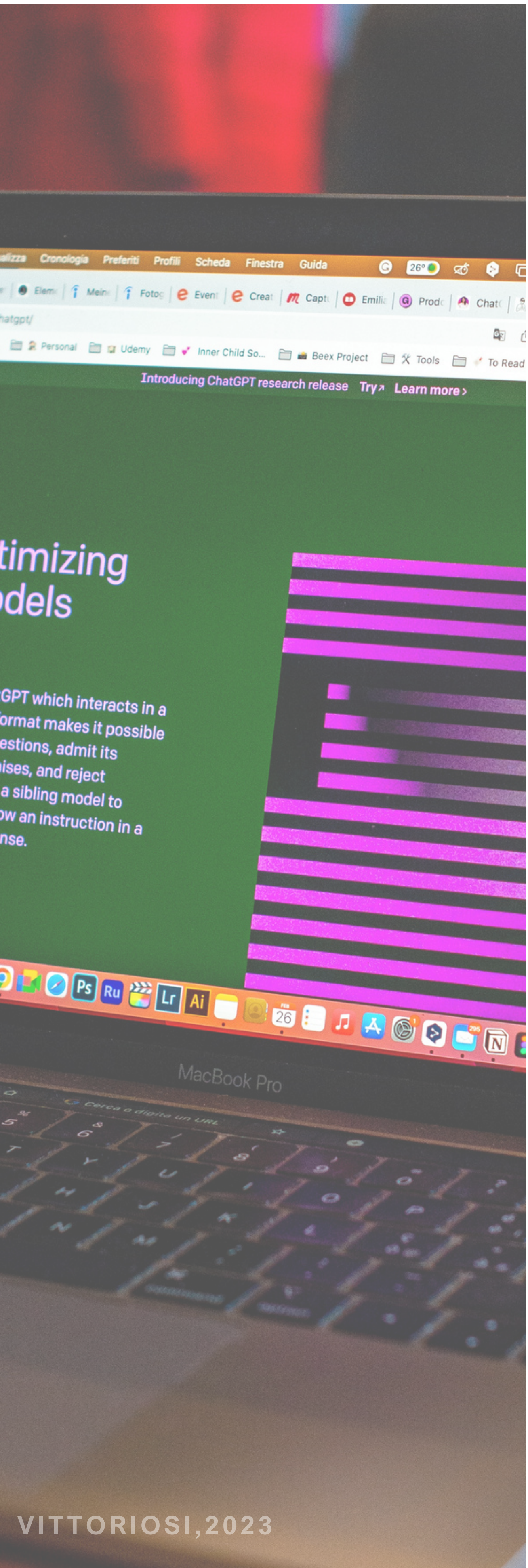
SHIELDING EU SECURITY: MITIGATING AI RISKS FOR STRATEGIC PROTECTION

CIVICA Capstone Project 2023: The Future of Europe

DISCLAIMER

Due to the timeframe of the course, the policy paper only reflects the political debate up to the 5th of December. Therefore, the recent trialogue negotiations between the 6th and 9th of December are not part of the analysis.

INTRODUCTION



VITTORIOSI,2023

In November 2022, OpenAI introduced ChatGPT. It gained an astonishing 100 million visitors in just two months (see Figure 1), 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 Artificial Intelligence (AI) can generate human-like text and create diverse multimodal content like images, video, and audio from various data sources, making these innovations possible.

However, generative AI can also produce 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). As a result, the European Union (EU) engages 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.

WHAT THE EU IS DOING

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 to safely utilise AI systems in the EU. 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 that could produce 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 disclosure of AI-generated or modified content. Since then, hundreds of amendments have been submitted, many of which relate to foundation models and generative AI, as they were barely mentioned in the initial drafts.

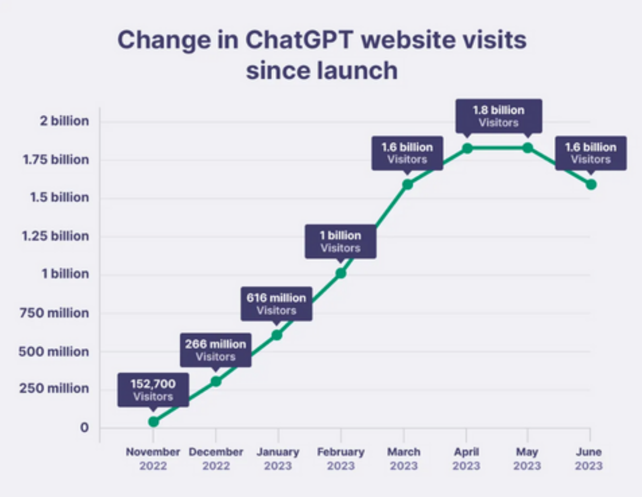


Figure 1: Evolution of the number of visitors of ChatGPT since its launch (Ver Meer, 2023)

In June 2023, the European Parliament voted for its negotiating position with more substantial and specific obligations for generative AI. These comprise the mandatory use of state-of-the-art safeguards against creating content that violates EU law, more substantial transparency obligations, and the public sharing of a summary of the copyrighted training data used. In addition, generative AI, categorised as a subset of foundational models, must also comply with the obligations imposed on foundation model providers. These include implementing data governance to ensure unbiased and appropriate datasets, demonstrating mitigation of reasonably foreseeable risks, and mandatory registration in an EU database. Part of this tiered approach is also compliance with the general guidelines for the risk category (Barani & Van Dyke, 2023) (refer to Figure 3).

As the European Parliament proposed amendments to the Commission's text, the legislation entered the trilogue phase — the inter-institutional negotiations between the Council of the EU and the European Parliament, supported by the European Commission. In the previous trilogue meetings, it seemed that a consensus could be found on the suggested stricter obligations, but only for the most powerful instruments and in a tiered manner. However, the negotiations have reached a standstill, as the powerful alliance around France, Germany, and Italy opposes any binding rules for foundation models and generative AI.

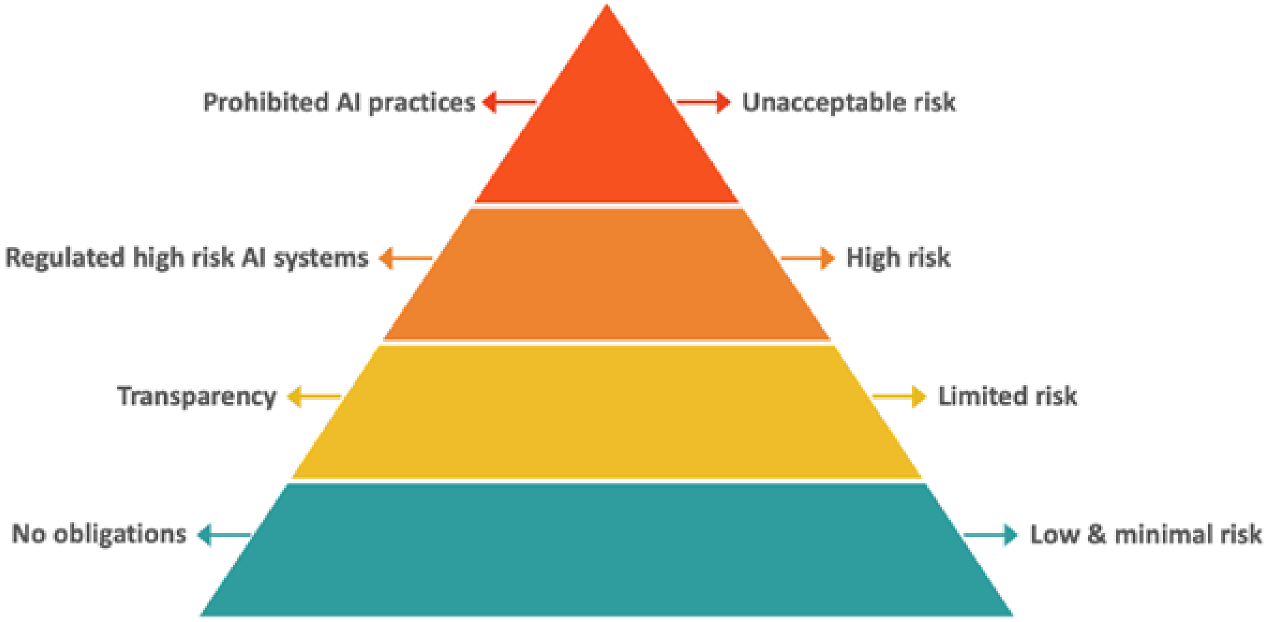


Figure 2: Different types of risks associated with AI systems as outlined in the initial AI Act by the European Commission (based on European Parliament, 2023)

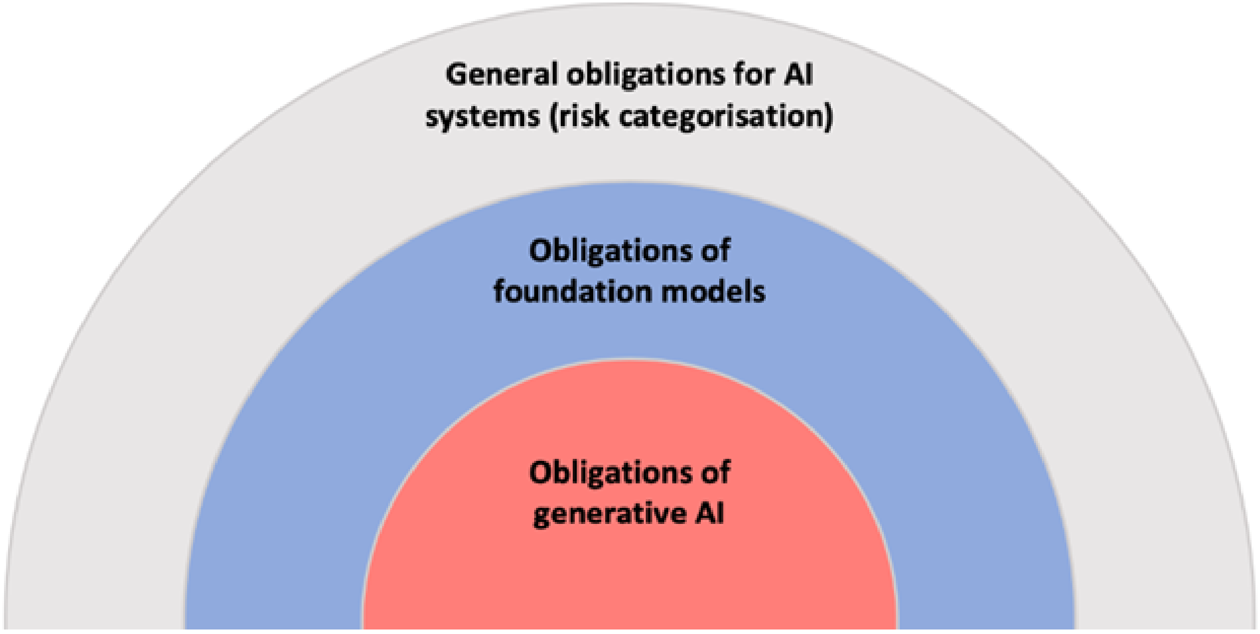

General obligations for AI systems (risk categorisation)
Obligations of foundation models
Obligations of generative AI

Figure 3: Tiered obligation structure for generative AI in the EU Parliaments proposal from the 14th of June 2023 (based on Barani & Van Dyke, 2023)

They are pushing for much looser rules, requiring only a self-regulatory code of conduct, called model cards, to ensure a minimum of transparency and safety. Penalties for violations are deliberately left out. This move is motivated by fear of being left behind in this forward-looking technology and the country's economic interests since the most promising AI companies, Mistral AI and Aleph Alpha, are based in France and Germany.

Overall, the policy development highlights the fundamental problem with regulating generative AI and foundational models. On the one hand, regulating these models and ensuring transparency and accountability cannot simply be done with the technology-neutral, risk-based approach to their applications since the latter are, by definition, unaware of these models. On the other hand, regulating the entire technology could, eventually, damage the economic strength of the EU. Thus, a solution can only be implemented if it represents a political compromise and internalises the apparent dilemma between economic interests and civil rights protection in the best feasible way. Failure to reach an agreement between the parties before the 2024 elections could lead to a worst-case scenario of 27 different regimes. This would jeopardise the basic idea of creating a harmonised legal framework in the European single market.

FLAWS OF AI REGULATION PROPOSALS

1. Alliance Proposal

The Alliance has proposed placing the responsibility of protecting civil rights in the hands of multinational technology companies. However, this decision has raised concerns about the lack of democratic accountability. Profit-driven business models may prioritize profits over civil rights protection, leading to conflicts of interest. A similar scenario happened with social media, threatening the democratic order in their quest for users' attention[1].

The unequal influence of these companies through lobbying and the use of significant resources is also problematic. Reconciling the different points of view and interests of companies from different countries can become a considerable issue when relying solely on self-regulation.

Additionally, the proposal does not impose any democratic or institutional enforcement obligations, which increases the likelihood of wrongdoing going unnoticed and unpunished. The lack of external or independent bodies to ensure compliance makes it challenging to hold individuals or organizations accountable for their actions. The lack of clear guidelines means that rules can be interpreted freely, leading to misconduct and rule-breaking. The lack of enforcement and accountability can also lead to greenwashing, perpetuating a cycle of misconduct and failure to meet ethical and legal standards.

[1] Specific threats are e.g., filter bubbles and echo chambers, fragmentation, polarisation, and disinformation (“fake news”) (Stark, Magin & Geiß, 2021).

Establishing a collaborative approach that incentivises private entities to contribute and align with social impact objectives is essential. However, relying solely on mutual trust and establishing a close working relationship between government officials and private companies, which is necessary for adequate data protection, may face insurmountable challenges due to conflicts of interest. Overall, this approach does not align with the EU's ambition to be the leading authority in safely regulating AI.

2. EU Parliament Proposal

The EU Parliament's proposal on AI regulation aims to protect society from the misuse of generative AI while leaving room for improvement. It proposes an ethical framework that prioritises fundamental rights and prevents privacy violations and discriminatory practices. It also adopts a risk-based approach that seeks to distinguish between several types of threats and takes appropriate measures. However, while the proposal provides for stricter mandatory transparency rules for foundation models, there are some concerns about the distribution of generated content. Misinformation and deep fakes generated by AI continue to proliferate, and it is rare for content to be promoted as AI-generated.

In addition, the rules could be burdensome for businesses and stifle innovation, especially for small and medium-sized enterprises (SMEs). The rules must not harm competition or the development of the AI landscape in the EU. There is also a risk that AI innovation will move out of the EU to regions with more lenient regulations. Ultimately, the aim should be to strike a more delicate balance between fostering innovation in AI and protecting the rights and interests of individuals and society.

WHAT THE EU SHOULD DO

In response to the previously identified issues, the following recommendations are proposed:

1. Implement a tiered Risk-Approach for Generative AI Models

In line with the European Parliament's proposal, explicit and strict obligations for generative AI providers are recommended due to the inherent risk of the models. However, in order not to stifle innovation for smaller AI companies, a tiered approach based on the size or power of the models is preferred. In this way, small AI developers are not disproportionately restricted compared to established technology companies. Furthermore, stricter rules for the big players would result in a downstream shift of power to users and application providers. Both groups would therefore face lower liability risks and compliance costs, which would encourage application innovation for SMEs (KIRA Center for AI Risks & Impacts, 2023).

2. Establish an AI Office for Cooperation and Innovation

Understanding the unique challenges and considerations of generative AI in the healthcare, gaming, and military sectors is critical. Given the rapid development of generative AI, regulations must be reviewed frequently. To accommodate emerging generative AI products that may not fit into existing categories, an independent European institution such as the proposed AI Office is needed (European Parliament, 2023, Amendment 85).

With the global AI market expected to reach $1.81 trillion by 2030 (Howarth, 2021), it is vital to foster innovation and engage with AI providers to continually develop regulations that work for everyone. In addition, encouraging diverse participation and knowledge sharing is key to achieving outcomes that benefit all stakeholders. The AI Office should also seek to collaborate with other economies to increase the impact of the act and create a global level playing field in the future.

3. Strengthen cooperation with business on Social Media Awareness

The AI Act still needs to address the challenges of identity theft through Deep Fakes and the spread of misinformation. Although AI-generated content should be labelled accordingly, its distribution is currently unregulated. Social media platforms have become the primary channels for spreading misinformation, as noted by Muhammed and Mathew (2022). To combat this, transparency guidelines regarding AI tools' visibility and mandatory labelling of AI content should extend to all generative AI tools, regardless of model size. Collaborative efforts between the EU and companies such as Facebook, TikTok, and YouTube are necessary to ensure the safe use of generative AI systems and social media by implementing real-time fact-checking and identifying AI content. Addressing the challenges of copyright infringement and deepfakes requires increased public awareness and education. Public education efforts should focus on educating individuals about the risks and realities associated with deepfake technology. Meanwhile, governments should provide accessible platforms for the public to report copyright infringement and deepfakes. Effective deepfake detection programs require collaboration between governments and different sectors, highlighting the importance of a central AI office.

REFERENCES

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. [Online]. 22 April. [Accessed 2 December 2023]. Available from: https://www.economist.com/science-and-technology/2023/04/19/how-generative-models-could-go-wrong?utm_medium=cpc.adword.pd&utm_source=google&ppccampaignID=18156330227&ppcadID=&utm_campaign=a.22brand_pmax&utm_content=conversion.direct-response.anonymous&gad_source=1&gclid=CjwKCAiAg9urBhB_EiwAgw88mU4UXWi5mRafeljjGEwyBmoujtzxs5ircbptGPckhmyhWXkQJiuX8BoCkDEQAvD_BwE&gclsrc=aw.ds.

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].