Simulating Policy Impacts: Developing a Generative Scenario Writing Method to Evaluate the Perceived Effects of Regulation

Julia Barnett[1], **Kimon Kieslich**[2], **Nicholas Diakopoulos**[1]
[1]Northwestern University,
[2]Institute for Information Law, University of Amsterdam
juliabarnett@u.northwestern.edu, k.kieslich@uva.nl, nad@northwestern.edu

Abstract

The rapid advancement of AI technologies yields numerous future impacts on individuals and society. Policymakers are tasked to react quickly and establish policies that mitigate those impacts. However, anticipating the effectiveness of policies is a difficult task, as some impacts might only be observable in the future and respective policies might not be applicable to the future development of AI. In this work we develop a method for using large language models (LLMs) to evaluate the efficacy of a given piece of policy at mitigating specified negative impacts. We do so by using GPT-4 to generate scenarios both pre- and post-introduction of policy and translating these vivid stories into metrics based on human perceptions of impacts. We leverage an already established taxonomy of impacts of generative AI in the media environment to generate a set of scenario pairs both mitigated and non-mitigated by the transparency policy in Article 50 of the EU AI Act. We then run a user study ($n = 234$) to evaluate these scenarios across four risk-assessment dimensions: severity, plausibility, magnitude, and specificity to vulnerable populations. We find that this transparency legislation is perceived to be effective at mitigating harms in areas such as labor and well-being, but largely ineffective in areas such as social cohesion and security. Through this case study we demonstrate the efficacy of our method as a tool to iterate on the effectiveness of policy for mitigating various negative impacts. We expect this method to be useful to researchers or other stakeholders who want to brainstorm the potential utility of different pieces of policy or other mitigation strategies.

Introduction

In addition to creating a suite of positive benefits, from facilitating ease of culturally situated natural language creation and translation (Yao et al. 2023), to assisting with medical imaging (Chen et al. 2022), and empowering accessible forms of education (Alasadi and Baiz 2023), generative AI has given reason for concern in many aspects of society. Any implementation of a system using generative AI, or any algorithmic tool for that matter, can have unintended downstream effects. It is essential to have a comprehensive understanding of these potential impacts when considering how to make decisions regarding the usage of these tools.

Experimenting with deployment of an algorithmic system and experiencing the tangible impacts can have costly effects on the individuals and society affected by it. An approach that can address this by helping to develop a comprehensive understanding of potential impacts prior to deployment is anticipatory ethics together with participatory foresight (Sarewitz 2011; Brey 2017). Anticipatory ethics guides scientific and technological advancement in a humane manner by using methodological approaches to consider potential impacts of the technology at all stages of development (Diakopoulos and Johnson 2021). Participatory foresight shifts this analysis to include the many stakeholders affected by the technology rather than simply the algorithm designer or model deployer (Barnett and Diakopoulos 2022; Bonaccorsi, Apreda, and Fantoni 2020). This, too, can be costly due to the involvement of a variety of stakeholders to get a comprehensive understanding of possible effects. In this work we begin to explore the potential for low cost large language models (LLMs) to contribute to this method.

Some LLMs, such as GPT-4, have been trained on more data than any human could consume in a millennium (Achiam et al. 2023). As such, they are able to learn enough about the world in terms of knowledge and relationships to be able to (to some extent) simulate some underlying cause and effect relationships in the world (Li, Nye, and Andreas 2021; Cai, Liu, and Song 2023). We aim to utilize this underlying 'world knowledge' in such models to simulate the impacts of AI in society given a policy implementation such as a particular Article of the EU AI Act (Commission 2024). We do so by utilizing written scenarios as small world representations that reflect and embed social and causal constraints within their text. Specifically, we prompt GPT-4 to write a large and varied set of scenarios each in terms of a particular impact about generative AI (e.g., sensationalism in regards to media quality impacts) to take place in the US in the next five years. We then introduce a mitigation strategy in terms of a policy implementation, and ask the model to re-write a given scenario in light of the introduced policy text. In this study we examine the impact of Article 50 of the EU AI Act as a specific expression of a set of policy ideas related to transparency obligations. To evaluate the quality of the scenarios and measure the influence of the policy on the impacts represented in the scenarios we then run a user study asking a set of human raters to read both the original

and re-written scenario and then rate them on four dimensions that are of high relevance in impact assessment (IA) and scenario-writing literature: severity, plausibility, magnitude, and specificity to vulnerable populations.

We find that the method we develop to generate written scenarios and re-write them under a policy condition is effective, producing scenarios that are largely considered as plausible by our study participants. Moreover, we were able to collect data to help understand the perceived severity, magnitude, and specificity to vulnerable populations of the various impacts that were represented by the scenarios generated. We show that the setup we develop here is able to demonstrate a difference in perceived impacts when the policy was simulated as part of the scenario, drawing attention to areas of impact where the policy may be more or less effective. These findings generally demonstrate that the approach developed for using LLMs to generate and re-write scenarios can be a valuable first step for stakeholders wishing to explore policy options for mitigating impacts, such as policy makers or researchers prior to more costly evaluation methods such as experiments, or pilot policy deployment. In our discussion we expand on the capabilities and limitations of this method and note the areas in which extra human evaluation and monitoring is needed. By demonstrating the viability of this method, we lay the groundwork to develop tools for policy makers and other non-technical stakeholders to more easily identify potential impacts and evaluate policy proposals as part of an anticipatory governance paradigm.

In sum, our work is primarily composed of the following two contributions: (1) we **develop an approach for the evaluation of policies that may mitigate societal harms**. We use scenarios written by an LLM to convey impacts and then further **use the LLM to simulate an alternative version of the scenario under a policy condition**; and (2) we **evaluate the approach with human raters** by using a case study about negative impacts in the domain of generative AI in the media/info ecosystem and in light of the EU AI Act.

Related Literature

We now expand on literature in anticipatory impact and governance in order to frame how our method situates within established work to comprehensively assess risk and impact of various AI systems. Then we explore existing work investigating the ability of large language models to learn underlying relationships of the world to assess how we can leverage these knowledge representations for the purposes of simulating the effects of policy on specified harms.

Anticipatory Impact and Governance

In order to guide the deployment of AI in a beneficial way for individuals and society, it can be helpful to look ahead and anticipate a wide range of potential impacts. Recently, the European Union articulated the need for anticipatory risk management in stating that reasonably foreseeable risks should be identified before the implementation of AI systems (Commission 2024). Inherent to this call is a need to engage in foresight of how technology implementation could impact individuals, but also society as a whole. Consequently, many researchers, NGOs, and companies have engaged in systematic impact or risk assessment activities to categorize and classify those impacts (Stahl et al. 2023).

The National Institute of Science and Technologies (NIST) introduced a structure to improve the trustworthiness of AI that follows a map, measure, and manage voluntary framework to deploy AI technologies (NIST 2023). They acknowledge the sociotechnical (Shelby et al. 2023) nature of these risks, and suggest the use of this framework to both initiate AI risk management and bolster existing systems. The idea is to (1) map out the potential harms of these AI systems; this is an active area of research identifying AI harms such as (Kieslich, Diakopoulos, and Helberger 2024; Kieslich, Helberger, and Diakopoulos 2024; Barnett 2023; Weidinger et al. 2023; Chanda and Banerjee 2022). The measuring step (2) involves methods to quantify these risks and enable assessment of whether we are mitigating them successfully. Finally, the managing step (3) involves integration of responsible policies to prevent these harms from manifesting. Our proposed method allows stakeholders to measure harms in an exploratory manner with the goal of enabling people in positions of power (e.g., legislators and government officials) to introduce policy to manage these harms.

The goal of algorithmic impact assessment (AIA) is related to the calls for anticipatory governance studies, namely to identify plausible impacts a technology might cause in an early development stage (Selbst 2021; Fuerth 2011; Guston 2014). In identifying potential detrimental impacts early, mitigation strategies can be developed beforehand so that the realization of those harms can be prevented (Selbst 2021; Guston 2014). Importantly, these impacts are not only of technical nature, but encompass the sociotechnical interplay of AI and people (Moss et al. 2021). Thus, anticipating impacts refers not only to the technological features of AI, but also brings the challenge of anticipating how different users will utilize it and how that will scale to societal impacts.

Anticipating these impacts is difficult due to the vast number of deployment settings and the inherent uncertainty of future developments (Nanayakkara, Diakopoulos, and Hullman 2020). Though scholars, NGOs, policy-makers, and companies have deployed various methods to conduct impact assessment, for instance with literature reviews (Weidinger et al. 2022; Shelby et al. 2023; Hoffmann and Frase 2023; Bird, Ungless, and Kasirzadeh 2023; Barnett 2023), the collection of expert opinions (Solaiman et al. 2023) or developer decisions (Buçinca et al. 2023). While these methods allow for a broad collection of potential impacts, they are top-down in a sense that they heavily rely on (domain) expert knowledge, but exclude the view of non-expert stakeholders. Multiple studies have already pointed out that expert-led anticipations are prone to be biased, especially towards what they desire the technology to be capable of (i.e., wishful thinking) (Bonaccorsi, Apreda, and Fantoni 2020; Brey 2017). Thus, researchers have called for the inclusion of the voices of laypersons that enrich the current landscape of expert-driven assessments in providing real-world imaginations of technology impact (Moss et al. 2021; Metcalf et al. 2021). This form of participatory foresight (Nikolova 2014) also helps in democratizing impact assessment.

Scenario writing is one method that can be an effective approach to stimulate future thinking of laypeople (Amer, Daim, and Jetter 2013; Amos-Binks, Dannenhauer, and Gilpin 2023; Börjeson et al. 2006; Burnam-Fink 2015; Selin 2006). The goal of scenario-writing is to create narratives about future impact of technology that vividly show how a specific technology impacts the life of story characters. Numerous plausible and potential futures emerge that can then be used as a conversation starter to develop mitigation strategies or to classify user-centric impacts (Amer, Daim, and Jetter 2013; Burnam-Fink 2015). The impact assessment literature has acknowledged the value of scenarios (Meßmer and Degeling 2023) and several studies have relied on laypeople to compose future impact scenarios (Diakopoulos and Johnson 2021; Kieslich, Diakopoulos, and Helberger 2024; Kieslich, Helberger, and Diakopoulos 2024), use scenarios as a conversation starter for debate about ethics (Das et al. 2024), or provide situations to collect judgements related to AI ethics (Awad et al. 2018).

In this study, we combine the potential of scenarios with LLMs and layperson judgements to help assess impacts. Concretely, we rely on the impact identification work of Kieslich, Diakopoulos, and Helberger (2024), who utilized scenario-writing to map potential impacts of generative AI in the news environment. In our study, we leverage this framework and combine it with the potential of LLMs to both synthesize written scenarios and, importantly, to re-write them under a policy condition, allowing us to collect and compare judgements of the scenarios by people.

LLMs as Knowledge Representation

Large language models such as GPT-4 (Achiam et al. 2023), Gemini (Team et al. 2023), and Claude 3 (Anthropic 2024) have been trained on enormous quantities of data from the internet, such as web documents, books, code, images, and audio. They utilize proprietary datasets as well as open-source web data such as the continuously growing Common Crawl, which contains more than 250 billion pages of over a decade's worth of textual content. While these models primarily learn how to linguistically represent the way we communicate, they also learn knowledge representations of how our world operates. OpenAI has touted the ability of GPT-4 to pass the American Bar Exam with flying colors (Achiam et al. 2023). Though this has received push back as not being as successful as originally reported (Martínez 2024), it has still passed well above the acceptance rate and been shown to outperform lawyers on tasks such as contract review (Martin et al. 2024).

In addition to performing well on reasoning tasks, LLMs also are capable of mapping relationships between entities. Park et al. (2023) developed a sandbox game environment similar to the video game *The Sims* which was powered entirely by LLMs. In this simulated environment, generative agents powered by the LLMs produced believable individual and social behaviors, and knew how to interact reflecting norms of society. Li, Nye, and Andreas (2021) also assert that LLMs' abilities to model language come not only from statistical modelling of corpora, but also dynamic representations of meaning learned from the training data.

Even further, there is evidence that LLMs have the ability to map causal relationships. Long et al. (2023b) explore the ability of GPT-3 to build causal graphs, determining they can be used to create Directed Acyclic Graphs (DAGs), though recommend accompaniment by expert knowledge. Long et al. (2023a) later evaluate the utility of using the LLM itself *as* the "imperfect expert" in providing knowledge about the causality of various relationships. Tu, Ma, and Zhang (2023) used ChatGPT (GPT-3) to explore the LLM's ability to identify causal relationships in Neuropathic Pain Diagnosis; they determined the model served as a good assistive tool for causal discovery, but still required expert direction. Others have asserted that LLMs can assist with the advancement of research in causality (Kıcıman et al. 2023). Jin et al. (2023) go beyond common sense causal understanding and build a dataset to evaluate LLMs' ability to perform formal causal and chain-of-thought reasoning; these tasks prove to be much more challenging for LLMs, though GPT-4 performs the best out-of-the-box with overall 70% accuracy, second only to their specifically tuned causal model. Finally, Cai, Liu, and Song (2023) demonstrate that when LLMs are provided with domain-specific knowledge, they can produce sound causal reasoning, and can maintain causal reasoning without this specific knowledge.

We leverage the various forms of knowledge that models such as GPT-4 encode in this study in order to be able to both generate scenarios reflective of specific impacts and to then re-write these scenarios given information about a policy. In this way we are able to simulate variations of scenarios which make use of the knowledge and understanding of connections and relationships encoded in the model.

Methodology

The goal of this work, again, is to understand to what extent LLMs (specifically GPT-4 in this study) are capable of generating complex scenarios about a given set of impacts, and their ability to then generate mitigated versions of those scenarios in light of a specific policy. To implement this (see Figure 1), we iterated on prompts to generate a dataset of scenarios reflecting impacts in the media environment due to generative AI. We reviewed these scenarios to ensure they were adequately representing the impacts for which we prompted. We then designed a survey to allow respondents to assess these scenarios across four dimensions that are relevant to risk assessment and AI ethics (severity, plausibility, magnitude, and specificity to vulnerable populations). Finally, we aggregated the results across these four dimensions to demonstrate how to effectively translate the rich textual scenarios into quantified evaluations.

Scenario Generation We used GPT-4[1] (Achiam et al. 2023) for this study since at the time of data collection it outperformed other proprietary and open LLMs on the ChatArena benchmark which evaluates models according to pairwise human comparisons (Chiang et al. 2024)[2]. We used OpenAI's API for chat completions to generate scenarios. To

[1]"gpt-4-turbo" available via the OpenAI API in Feb. 2024.

[2]https://chat.lmsys.org/?leaderboard

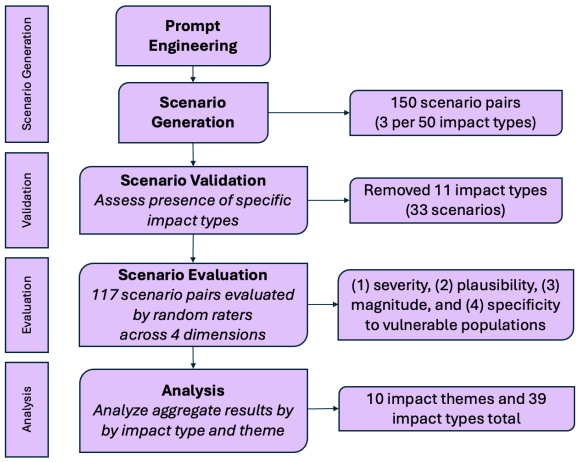

Figure 1: Flow diagram of study, starting with prompt engineering to generate 150 scenario pairs (3 scenarios for each of 50 impact types), then manual author validation check removed 33 scenario pairs (11 impact types), then had human raters evaluate 117 scenarios, and finally analyzed the aggregated data across impact themes.

generate scenarios illustrating various impacts we ground the approach with a set of 50 *impact types*, organized into 10 higher-level *impact themes* that describe various impacts of generative AI in the media and information ecosystem (Kieslich, Diakopoulos, and Helberger 2024). This "Kieslich taxonomy" as we refer to it here, was itself developed by analyzing scenarios written by a diverse array of respondents to an anticipatory scenario writing task. The 10 impact themes identified include: autonomy, education, labor, legal rights, media quality, political, security, social cohesion, trustworthiness, and well-being, and within these there are 50 specific impact types which we explicitly use in our scenario generation process and refer to in our findings. By leveraging an already established taxonomy of specific impacts, we not only ground the approach and addresses plausibility and bias issues created by more speculative LLM use, but also have an established thematic grouping to guide our analysis and interpretation. Future work may consider extending the methods developed here to other impact typologies and domains.

Developing a Prompting Approach. We initially experimented with simply prompting the model with the 10 general impact themes in the Kieslich taxonomy in order to understand the diversity with which GPT-4 could uncover potential impacts when only prompted with a general impact theme. We thus prompted GPT-4 to generate scenarios for each of the ten impact themes ten times each. In the Appendix[3], Figure A5 displays the number of specific impact types that were uncovered by the LLM, as well as how often various specific impact types were discussed.

When using this prompting approach the model was not able to generate scenarios reflecting each of the specific impact types outlined in the Kieslich taxonomy except for *trustworthiness*. More importantly, for each of the impact themes, the model tended to heavily skew toward discussing a couple of specific impact types and rarely touched on others. For instance, prompting only with the impact theme of *well-being* produced scenarios covering three out of four specific impact types from the typology, and generated scenarios discussed *mental harm* 90% of the time, never mentioned *physical harm*, and only discussed *addiction* and *reputation* two out of ten times each.

Due to this skew and non-comprehensive spread of impact types present in the scenarios when prompting only with impact theme, we moved on to experiment with prompting the model explicitly for each of the 50 impact types rather than at the level of the ten impact themes. It became evident that utilizing LLMs to generate scenarios was more effective when done with a human-in-the-loop providing contextual knowledge (in this case, the full Kieslich taxonomy) in order to guide creation of specific outputs—if we only used general impact themes we would not uncover as many impacts or the same level of detail. Thus, for the remainder of this paper, we evaluate scenarios generated by GPT-4 that were generated with inclusion of each specific *impact type*.

Creating Stimulus Material for the Study. For each specific *impact type*, we generated scenario pairs consisting of one illustrative version (S) to narrativize the impact and a corresponding policy-mitigated version (S') which was rewritten in light of a policy condition. We chose to generate 3 scenario pairs for each *impact type* in order to evaluate the reliability of the model for generating scenarios relevant to the impact while also providing some variance in the sample. This process resulted in the production of 150 scenario pairs. In order to remain consistent across all *impact types*, the only variation in the prompt was switching out the words for the *impact type* using the format "$\langle$*specific impact type*$\rangle$ in regard to $\langle$*impact theme*$\rangle$" (e.g., "*sensationalism* in regards to *media quality*" or "*polarization* in regard to *social cohesion*"); all other aspects of the prompt remained the same. The full prompts can be found in the Appendix A.2.

We refer to the prompt for the illustrative scenario, S, as Prompt 1, and the prompt for the policy-mitigated scenario, S', as Prompt 2.[4] For Prompt 1, We first included context to define what a scenario was, name the specific *impact type* in the context of the *impact theme*, and define generative AI. We then directed the model to write a short fictional scenario set in the United States in the year 2029 about this impact, and included specific instructions for the model not to introduce reflection or meta analysis in the scenario since we reasoned that this might influence or bias the interpretation of participants in our study. For Prompt 2, we indicated to the model that a piece of legislation was enacted and that it should rewrite the scenario in light of that legislation. For this study, we provided it the exact text of the final version of Article 50 of the EU AI Act (Commission 2024), which concerns transparency obligations for certain AI systems, including for general purpose AI models.

Scenario Validation. To assess the extent to which GPT-4 adequately generated scenarios for each of the impact types, we had two authors independently validate each of the 150

[3]Accompanying appendix found at: arxiv.org/abs/2405.09679.

[4]For each scenario pair we used a fresh state with temp. 0.7.

scenarios in order to assess whether the scenarios actually depicted the designated impact provided in the corresponding prompt. Evaluators rated each scenario as (1) depicting the impact, (2) somewhat depicting the impact, and (3) not depicting the impact. A scenario pair needed either both evaluators to assess it as depicting the impact or one evaluator saying it depicted and the other saying it was somewhat present to be considered valid. If zero or only one of the three scenarios were deemed valid, the specific impact type was removed from further evaluation. If two of the three scenarios passed the validation check, the third scenario was regenerated and then reassessed by both evaluators to ensure all three scenarios depicted the designated impacts.

After this process, we ended up with 39 of 50 of the original *impact types* and thus 117 scenario pairs for evaluation in our user study.

Study Participants For this study we recruited raters for the scenarios on Prolific that had an approval rating of at least 95/100, and 100 approved tasks. To ensure evaluators were familiar with the context of the scenarios (which all took place in the United States), we utilized evaluators who were proficient in English and resided the United States.

We excluded respondents that failed at least two attention checks. Our attention checks took the form of nonsensical statements for which there was a logical and common sense response (e.g., To what extent do you agree with this statement: "I've developed a way to photosynthesize, so I sustain myself on sunlight alone, like a plant."? This is to check your attention.)[5]. Evaluators were not allowed to take the survey multiple times. Fifteen raters were removed via attention checks (6% of 249 recruited) resulting in a final participant pool of 234 raters. 12% of evaluators did not consent to demographic data collection through Prolific, but the remaining sample was comprised as follows: a fairly equal dispersion of male and female identifying evaluators (51% F; 48% M; 1% prefer not to say), predominantly white (68%), followed by 11% Mixed, 8% Black, 7% Asian, and 5% other; and an average age of 38 (minimum: 19, maximum: 76).

We had six people evaluate each scenario pair in order to establish an average response across multiple independent raters who might each bring their own subjectivities to their ratings. We paid each evaluator to annotate a random set of three scenario pairs, which took approximately 14 minutes (final study median). We paid $3.25 for each completed set (an estimated pay of $14 per hour).

Survey Design We chose to assess the impacts portrayed in the scenarios across four dimensions: *severity*, *plausibility*, *magnitude*, and *specificity to vulnerable populations*.

The EU AI Act (Commission 2024) defines risk as "the combination of the probability of an occurrence of harm and the severity of that harm" positioning both prevalence and severity as key dimensions that need to inform a risk-based approach to addressing AI impacts in society. Assessing the severity and prevalence (what we term magnitude in this study) of impacts presented in scenarios is thus a crucial step to establishing mitigation priorities for AI systems, such as general purpose models that can cause a wide range of different harms (Meßmer and Degeling 2023).

Plausibility is an essential evaluation dimension for risk analysis (Glette-Iversen, Aven, and Flage 2022) and scenario quality (Amer, Daim, and Jetter 2013). Following the typology consolidated by Börjeson et al. (2006), we are generating *explorative* scenarios in this work; we need a measure to assess the degree to which the events in a scenario are realistic or consistent with known social dynamics such that one could reasonably conclude that a scenario *could* happen based on their world knowledge (Urueña 2019).

Finally, assessing specificity to vulnerable populations is not an established metric for evaluating scenarios, but it is widely understood that there is often a disparate impact on people not belonging to the majority power group when it comes to societal harms, including AI harms which can be magnified for vulnerable and minority groups (Benjamin 2019; Hagerty and Rubinov 2019; Shelby et al. 2023).

In our study we presented evaluators with a scenario (either S or S'), and asked them to answer the following questions with respect to the scenario they had just read:

1. "**Severity**" is defined as "the condition of being very bad, serious, unpleasant, or harsh". Please rate the **severity** of this scenario on a scale of 1 (not severe) to 5 (extremely severe).
2. "**Plausibility**" is defined as being reasonable to conclude something may happen. Please rate the **plausibility** of this scenario on a scale of 1 (not plausible) to 5 (extremely plausible).
3. "**Magnitude**" is defined as "great size or extent." Please rate the **magnitude** by considering the following: This scenario is an example that **could happen** to any number of people in the world. **How many people** do you think could be affected by the harms presented in this scenario? 1 (a small number of people); 5 (the majority of people in society).
4. "**Vulnerable Populations** are defined as "individuals who are at greater risk of poor physical and social health status". Please rate how **specific** you think the risk(s) presented in this scenario are to **vulnerable populations in comparison to non-vulnerable populations**. 1 (not specific to vulnerable populations); 5 (extremely specific to vulnerable populations).

We wanted to allow evaluators' intuition to guide them to evaluate these rather than have any sort of anchoring guide such as giving an explicit example of what would constitute 'extremely severe'. The definitions of the items were developed from dictionary definitions to elicit ratings from colloquial understandings rather than injecting expert definitions and biases into the evaluations. After a rater evaluated the first scenario, they were then presented with "a re-written version of the above scenario" (either S or S', whichever they did not see first) and performed the same evaluation on this second scenario. To mitigate the potential for ordering effects we randomly counterbalance the order of presentation: three evaluators saw each scenario pair in the order S,

[5]See: Prolific's Attention and Comprehension Check Policy. https://tinyurl.com/prolific-attention ; Our attention checks are available at: https://tinyurl.com/attention-checks

S', and three people saw the scenario in the order S', S. We later performed a paired sample t-test to assess if there were any ordering biases and found none.

Analysis and Results

We now evaluate the efficacy of GPT-4 as a scenario generator and iteration tool for simulating potential policy mitigation by using a case study in the domain of AI in the media ecosystem with Article 50 of the EU AI Act (Commission 2024). As detailed above in the methodology, we conducted a user study to evaluate GPT-4 generated scenario pairs comprised of: (S) a scenario illustrating a specified negative impact in the media environment and (S') a re-written version of that scenario that assumes the transparency legislation described in Article 50 of the EU AI Act was enacted.

For each impact theme below, we will examine (a) which scenarios GPT-4 was adequately able to generate when specifically prompted, (b) the human user study raw scores to understand how these impacts were perceived, and (c) the delta between S and S' to evaluate whether humans perceived the policy to be effective at mitigating the designated harm. We begin by examining overall significant trends, then proceed alphabetically by impact theme. Significance refers to results from paired sample t-tests, and we report significance at levels of $p < 0.05, 0.01$, and 0.001. All references to a taxonomy of impacts of generative AI in the media environment uncovered by human authors refer to the Kieslich taxonomy (Kieslich, Diakopoulos, and Helberger 2024).

Overall (*Figure 2; Table A1*) We first describe some high level findings from the user study. Of the 50 specific *impact types* from the established taxonomy of generative AI impacts in the media environment, GPT-4 was able to adequately depict 39 impacts when prompted specifically. This indicates that there are some areas where the prompting approach taken was not successful, and future work could explore prompts that define impact types more explicitly.

Evaluators generally thought the impacts depicted in these scenarios were relatively severe ($M_S = 3.66; SD_S = 1.08$), but mitigated in part by the introduced legislation ($M_\Delta = -0.38$***; $SD_\Delta = 1.01; p < 0.001$), with 40% of scenarios demonstrating a lower severity after the transparency legislation was introduced. They found them to be quite plausible ($M_S = 3.94; SD_S = 0.96$), with the legislation doing little to affect the plausibility ($M_\Delta = -0.03$*; $SD_\Delta = 0.84; p < 0.05$). Evaluators believed that the original scenarios had impacts affecting a greater number of people ($M_S = 3.79; SD_S = 1.00$) than the policy mitigated scenarios ($M_\Delta = -0.23$***; $SD_\Delta = 0.90; p < 0.001$), indicating they thought that when transparency legislation was introduced the harm affected a smaller amount of people. Finally the human evaluators at large had fairly mid-range assessments of specificity to vulnerable populations ($M_S = 3.06; SD_S = 1.24$), indicating that though these were slightly more impactful to vulnerable populations, they were still having an impact on the majority of society. This perception also lessened with introduced policy ($M_\Delta = -0.13$***; $SD_\Delta = 0.80; p < 0.001$), though this was a relatively smaller change in comparison to severity and magnitude.

For each of the four dimensions (severity, plausibility, magnitude, and specificity to vulnerable populations), the changes within the overall comparison of M_S and $M_{S'}$ were all significant at $p < 0.001$. In Figure 2 we display the significance of all impact themes. Within severity there were several statistically significant results, on which we elaborate below. None of the changes at the impact theme level in plausibility were statistically significant, which speaks to the perceived possibility of both versions of the scenario transpiring in the near future. Only *labor*, *media quality*, and *trustworthiness* had statistically significant changes in magnitude, and statistically significant changes in perceived specificity to vulnerable populations were only present in impacts regarding *political* and *well-being*.

Autonomy (*Table A2*) From the taxonomy of generative AI impacts, there were three main *impact types* within autonomy: *loss of control, loss of orientation*, and *machine autonomy*. Using tailored prompting, we were only able to adequately generate relevant scenarios for *loss of control* and *loss of orientation*. The scenarios all discussed the impacts in the context of information consumers, with *loss of orientation* also focusing on the impact on journalists. Overall, evaluators perceived the transparency legislation to be effective at lessening the severity of these autonomy impacts ($M_\Delta = -0.72$*; $SD_\Delta = 1.15; p < 0.05$) and thought they were highly plausible ($M_S = 3.94; SD_S = 0.96$).

Evaluators thought the scenarios depicting *loss of control* were much more severe ($M_S = 4.22$) than those depicting *loss of orientation* ($M_S = 3.33$), with autonomy impacts having an average severity score of $M_S = 3.78$. However, *loss of control* scenarios depicted impacts perceived to be heavily mitigated by transparency legislation: these rewritten scenarios had on average $M_\Delta = -1.00$, one of the greatest deltas among all *impact types*. For instance, in one of the scenario pairs a reader named Mary felt out of control in her own ability to leave the highly partisan echochamber of political information encompassed by each news story she encountered. However, the legislation mandating that each AI-story be marked as "AI-generated" allowed her to question the content and intentionally find some human-authored articles with opposing arguments, ultimately leaving her in charge of her own opinions. Evaluators on average rated this as much less severe than the non-mitigated scenario ($M_S = 4.67; SD_S = 0.47; M_\Delta = -2.00; SD_\Delta = 0.82$).

Evaluators believed these issues affected a large portion of society ($M_S = 4.06; SD_S = 0.97$) with the second highest average scores for magnitude. They believe the mitigated impacts affected slightly fewer people ($M_\Delta = -0.17; SD_\Delta = 1.01$), and made the impacts less specific to vulnerable populations ($M_\Delta = -0.44; SD_\Delta = 0.83$).

Education (*Table A3*) Though when generally prompted for education impacts GPT-4 struggled to produce relevant scenarios (20% of the time; Figure A5), it was adequately able to generate relevant scenarios when prompted specifically for *critical engagement* and *literacy* impacts. One scenario each had to be regenerated for these specific impacts—

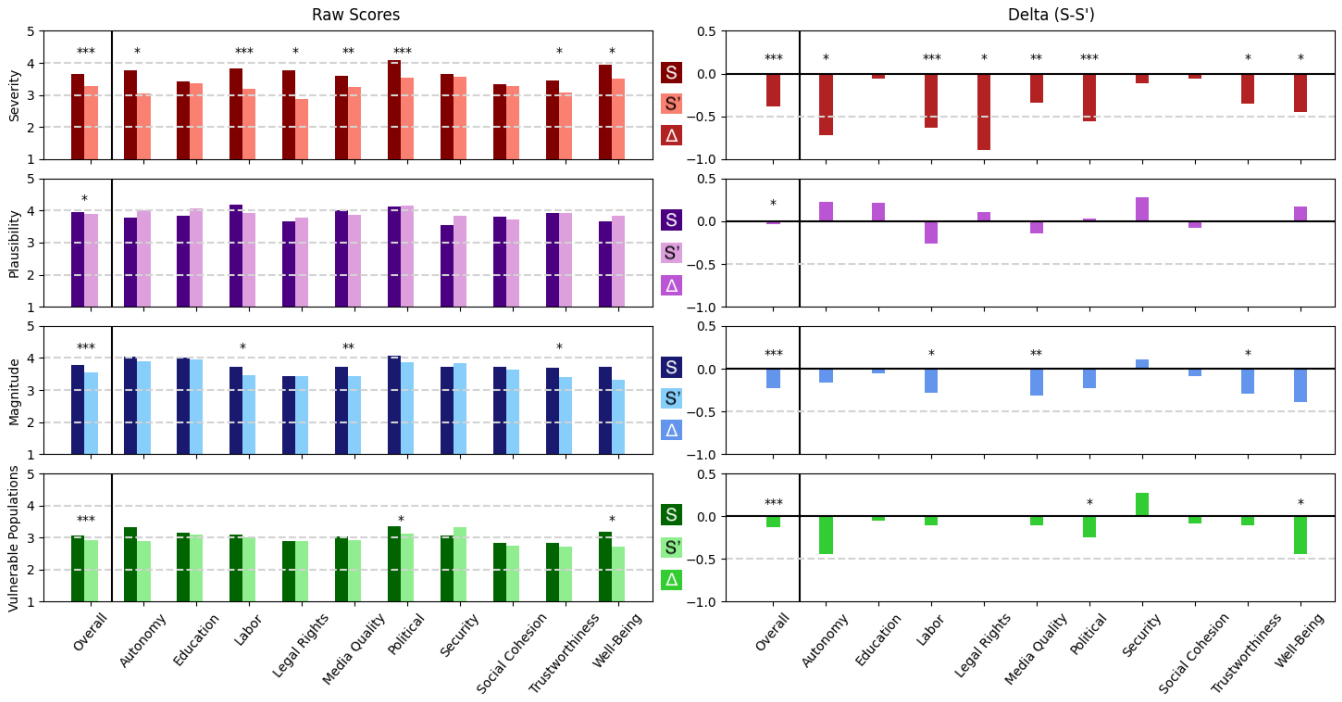

Figure 2: Bar chart displaying from top to bottom the four dimensions: severity, plausibility, magnitude, and specificity to vulnerable populations. Left plot: we first display the mean raw scores (M_S and $M_{S'}$) for each impact theme (in alphabetical order: overall, autonomy, education, labor, legal rights, media quality, political, security, social cohesion, trustworthiness, and well-being). Right plot: we display the deltas ($M_S - M_{S'}$) for each theme. For significance levels we include above the bars: * for $p \leq 0.05$, ** for $p \leq 0.01$, and *** for $p \leq 0.001$.

these two scenarios both discussed fake news about schools, not about actual education impacts. The 6 scenario pairs within education all focused on readers and consumers of news media, typically within a K-12 school environment. None of the changes between the illustrative and policy-mitigated scenarios for education impacts were significant.

Labor (*Table A4*) There are five specific impact types within the labor theme: *changing job roles, competition, job loss, loss of revenue,* and *unemployment*. Of these, the only specific *impact type* for which it was difficult to successfully generate scenarios was *loss of revenue*—this was the only *impact type* which took many additional attempts to generate relevant scenarios. All labor scenarios almost exclusively focused on the lives of journalists, however some scenarios detailing *changing job roles* described the effects on readers and consumers of news media, and *loss of revenue* detailed the effects on news organizations as a whole.

On average, labor impacts were seen as the third most severe (only less than political and well-being impacts). Even further, the transparency legislation was perceived as having a statistically significant effect on the severity of these impacts ($M_\Delta = -0.63$***; $SD_\Delta = 1.11; p < 0.001$). Similarly, respondents perceived the mitigated scenarios as impacting a smaller subset of society ($M_\Delta = -0.28$*; $SD_\Delta = 0.80; p < 0.05$).

The most severe of the *impact types* were *changing job roles* and *unemployment*, with severity scores of $M_S = 4.30$ ($SD_S = 0.90$) and $M_S = 4.22$ ($SD_S = 0.63$), respectively. However, transparency legislation seemed to have a large impact on *changing job roles*, *job loss*, and *unemployment* ($M_\Delta = -1.1$*, -0.78, and -0.56), indicating that transparency could potentially lessen the perception of impacts in these areas. Labor impacts were seen as the most plausible ($M_S = 4.17; SD_S = 0.79$) on average of all impact themes, however the policy-mitigated scenarios were perceived as less plausible ($M_{S'} = -0.26; SD_{S'} = 0.88$).

Legal Rights (*Table A5*) Perhaps the most elusive of all impact themes, we were only able to adequately generate relevant scenarios for one of four impact types for legal rights: *copyright issues*. *Freedom of expression, lack of regulation,* and *legal actions* all produced scenarios detailing other types of impacts. Thus, when we discuss the impacts of legal issues we are exclusively discussing *copyright issues*. These scenarios all described AI news generators unknowingly stealing or leaking copyrighted content unbeknownst to journalists or the creators of said content.

Legal rights impacts, though relatively severe ($M_S = 3.78; SD_S = 0.63$), had the largest perceived decrease in severity due to transparency legislation ($M_\Delta = -0.89$*; $SD_\Delta = 0.87; p < 0.05$). They were seen as reasonably plausible ($M_S = 3.67; SD_S = 0.47$), with a slight increase in plausibility when legislation mitigated the harm

($M_\Delta = 0.11; SD_\Delta = 0.74$). They seemed to affect a moderate portion of society ($M_S = 3.44; SD_S = 0.68$) and with moderate to low specificity to vulnerable populations ($M_S = 2.89; SD_S = 0.99$) with no change to magnitude or specificity to vulnerable populations when the transparency legislation was introduced.

Media Quality (*Table A6*) The most prolific of all impact themes, there were 16 original impact types within media quality of which we were able to prompt GPT-4 to generate 12: *accuracy/errors, clickbait, credibility/authenticity, distinction between journalism and ads, ethics, journalistic integrity, lack of diversity/bias, lack of fact-checking, loss of human touch, over-personalization, sensationalism*, and *superficiality*. We were unable to generate relevant scenarios for *accountability, attribution, explainability,* and *reframing*, which we hypothesize happened due to the intangibility and relative abstraction of these impact types. Typically in these instances the scenarios reverted to discussing *fake news/misinformation* and *accuracy/biases*—impacts we specifically prompted for elsewhere. The media quality scenarios focused mostly on digital news outlets, with the generative AI typically taking the form of a news generator. The impacts focused on both journalists and consumers of news media.

Transparency legislation was perceived as having a statistically significant effect on the severity of these impacts ($M_\Delta = -0.34$**; $SD_\Delta = 1.16; p < 0.01$), as well as affecting fewer people when mitigated by the transparency legislation ($M_\Delta = -0.31$**; $SD_\Delta = 1.05; p < 0.01$). The transparency legislation can therefore be seen as particularly relevant to impacts relating to this theme.

Within media quality, the specific impact types with the highest severity were *sensationalism, clickbait, credibility/authenticity*, and *lack of fact-checking* with scores of $M_S = 4.38, 4.00, 3.90$, and 3.78, respectively. The *impact types* for which the transparency legislation had the strongest mitigation effect were *journalistic integrity, clickbait, ethics*, and *lack of fact-checking*, with $M_\Delta = -0.89$*, $-0.88, -0.88$, and -0.78, respectively. On average media quality impacts were perceived as very plausible ($M_S = 4.00; M_{SD} = 0.98$), with legislation mitigation only slightly affecting the plausibility ($M_\Delta = -0.14; SD_\Delta = 0.84$). The specific impact types that were perceived as impacting a relatively substantially large population of people were *clickbait, sensationalism*, and *lack of diversity/bias* ($M_S = 4.38, 4.25$ and 4.10, respectively).

Political (*Table A7*) Each of the specific impact types within political impacts were successfully generated by GPT-4: *fake news/misinformation, manipulation, opinion monopoly*, and *political consequences*. This was the impact theme that specifically affected public figures (mainly politicians) more so than any other impact theme, but the scenarios also focused on the effects on society.

Political impacts across the board tended to be perceived as the most severe of all impact themes ($M_S = 4.08; SD_S = 0.76$), with *fake news/misinformation* having the highest severity ratings of all *impact types* with a score of $M_S = 4.56$ ($SD_S = 0.50$), and *manipulation* not trailing too far behind ($M_S = 4.22; SD_S = 0.92$). On average, transparency legislation seemed to mitigate the severity ($M_\Delta = -0.56$***; $SD_\Delta = 0.86; p < 0.001$) of political impacts and even more so in *fake news/misinformation* ($M_\Delta = -0.78$*; $SD_\Delta = 0.63; p < 0.05$) and *manipulation* ($M_\Delta = -0.67$*; $SD_\Delta = 0.82; p < 0.05$). One scenario portraying *fake news* detailed a fabricated political scandal concocted by an AI news generator and widely disseminated to the public that ultimately led the political candidate to drop out of the race. In the transparency mitigated scenario, an "AI-generated" tag was attached to the piece, and news consumers all paid attention to the human-verified fact checking of the inaccuracies of the purported scandal.

Scenarios detailing political impacts were seen as very plausible ($M_S = 4.11; SD_S = 0.97$), second only to labor. They also notably were one of the only impact themes for which the transparency legislation had a perceived significant impact on the specificity to vulnerable populations ($M_\Delta = -0.25$*; $SD_\Delta = 0.68; p < 0.05$), indicating that the perceived effect of this legislation may be even more beneficial to those more disparately affected by AI harms.

Security (*Table A8*) Though GPT-4 only adequately detailed security impacts half the time when prompted generally with the general security impact theme—and only *hacking*, at that—it was able to consistently generate scenarios for *cybersecurity* and *hacking* when specifically prompted. These scenarios tended to focus on public figures (politicians, celebrities, business people, etc.) as victims of attacks and public perception as a result of *hacking* and *cybersecurity* breaches. There were no significant perceived changes among any of the four dimensions in security impacts.

Social Cohesion (*Table A9*) Of all the specific impact types within social cohesion, we were able to successfully prompt GPT-4 to generate scenarios for four out of five: *dissatisfaction, polarization, real-world conflicts*, and *social divide*. Interestingly, the only impact type we were unable to generate scenarios for was *discrimination*—this should be a straightforward generation and we hypothesize this could have been an issue with content moderation preventing these topics from being discussed in the scenarios. Social cohesion scenarios focused more on the readers than any other category—all other impact themes had at least a minor focus on journalists or public figures. All focused on some degree of *polarization*, even when that was not the specific prompt.

On average, impacts within social cohesion tended to be less severe than other impacts detailed in the taxonomy ($M_S = 3.33; SD_S = 0.94$), with the exception of *polarization* ($M_S = 3.78; SD_S = 1.03$). There were no significant perceived changes within social cohesion impacts.

Trustworthiness (*Table A10*) We were able to generate adequately relevant scenarios for each of the *impact types* within trustworthiness: *discernment of fact and fiction, information chaos, media fatigue, mistrust*, and *overreliance on AI*. These scenarios tended to focus on the relationship between readers and journalists or readers and the news environment at large. They also feature a trajectory towards a lack of trust between society and the news we consume.

Though comparatively less severe than other impact

themes pre-policy mitigation ($M_S = 3.44; SD_S = 1.13$), legislation had a statistically significant perceived effect on lessening this ($M_\Delta = -0.36$*; $SD_\Delta = 0.90; p < 0.05$), and similarly lessened the perceived amount of people impacted ($M_\Delta = -0.29$*; $SD_\Delta = 0.81; p < 0.05$).

All specific impact types within trustworthiness had extremely similar scores, with the exception of *overreliance on AI* affecting the most people ($M_S = 4.33; SD_S = 0.47$), and *media fatigue* tending towards being perceived as more severe when transparency legislation was introduced ($M_\Delta = +0.22; SD_\Delta = 0.63$). Intuitively this makes sense, since labeling media as AI-generated creates another layer of information for people to process and make sense of and which therefore may contribute to fatigue.

Well-Being (*Table A11*) Of the four specific *impact types* identified by the taxonomy for well-being, GPT-4 only adequately generated relevant scenarios for *addiction* and *mental harm*. It was unable to generate scenarios when prompted specifically for *reputation* impacts, which is interesting given that this harm was present in other scenarios, such as when prompting for political and media quality impacts. Each time we prompted the LLM to generate scenarios for *physical harm*, it defaulted to discussing *mental harms* such as anxiety and paranoia, which can manifest physically but we classified as *mental harms* for the purpose of this study. Scenarios involving *addiction* focused on the addictive nature of generative media due to extreme personalization and proliferation. *Mental harms* as noted focused on the resulting anxiety and other mental health concerns arising from generated content, constant deluge of news media (whether fact or fiction), and impact of fake news and misinformation on the psyche.

Well-being impacts had the second highest severity ranking ($M_S = 3.94; SD_S = 0.91$), second only to political harms. Transparency legislation was also perceived to be effective at reducing this severity ($M_\Delta = -0.44$*; $SD_\Delta = 0.76; p < 0.05$). *Addiction* was perceived as slightly more severe than *mental harm* ($M_S = 4.11; SD_S = 0.87$ vs. $M_S = 3.78; SD_S = 0.92$, respectively), though the transparency legislation had a tendency to mitigate both even if this pattern did not reach statistical significance ($M_\Delta = -0.67; SD_\Delta = 0.82$ and $M_\Delta = -0.22; SD_\Delta = 0.63$, respectively). Legislation in this impact area also seemed to have a significant effect on reducing the specificity to vulnerable populations ($M_\Delta = -0.44$*; $SD_\Delta = 0.76; p < 0.05$).

Discussion and Future Work

In this work we developed a method utilizing LLMs to generate scenarios to evaluate policies that may mitigate negative impacts on society. We demonstrate this method by simulating scenarios both mitigated and non-mitigated by transparency legislation (Article 50 of the EU AI Act) to convey negative impacts of generative AI in the media ecosystem. We then asked human evaluators to rate these simulated futures in order to gauge human perception of the severity, plausibility, magnitude, and specificity to vulnerable populations of this policy mitigation. This method offers an opportunity to conduct policy evaluation on a given set of impact types in a straightforward manner. We now recap our high level findings, discuss limitations and capabilities of this method, elaborate on immediate possibilities of future research, and situate this method in risk-assessment work.

This study follows the goal of algorithmic impact assessment studies as it aims to map negative impacts of a given technology and explore mitigation strategies in how to prevent them (Selbst 2021; Moss et al. 2021). Using scenario-writing as a bottom-up approach of illustrating detrimental impacts of AI technology (Amer, Daim, and Jetter 2013; Meßmer and Degeling 2023), we supplement the top-down driven landscape of AIA. In this study, we introduced LLMs as a cost-effective way to create scenarios based on an already established impact typology (Kieslich, Diakopoulos, and Helberger 2024), evaluating scenarios across 10 impact themes relevant to the impact of generative AI on the media and information ecosystem: *autonomy, education, labor, legal rights, media quality, political, security, social cohesion, trustworthiness*, and *well-being*. Plausibility scores for the scenarios we generated remained relatively high across the board, indicating that GPT-4 was able to produce scenarios of impacts that were perceived to be genuinely possible depictions of the near-future impact of generative AI in the information ecosystem. The introduced transparency legislation was perceived to mitigate depicted negative impacts (whether in terms of severity, magnitude, or specificity to vulnerable populations) with statistical significance for *autonomy, labor, legal rights, media quality, political harms, trustworthiness*, and *well-being*. It did not seem to be impactful in the areas of *education, security*, or *social cohesion*, suggesting that perhaps other policy interventions are needed to combat negative impacts in these areas. For one specific impact, *media fatigue*, data suggests that the policy could even increase the severity of the impact.

Our findings demonstrate that LLMs can sufficiently simulate narratives about AI's impact and mitigation strategies, further contributing to the ongoing exploration of the ability of LLMs to map underlying complex relationships of the world, adding to discourse amongst Park et al. (2023); Tu, Ma, and Zhang (2023); Kıcıman et al. (2023); Cai, Liu, and Song (2023). By conducting a user study we are able to understand how individuals perceive the potential impacts of given policies—a proof-of-concept for a future more democratized process that would leverage a representative sample. Such input could inform a wider matrix of trade offs that policymakers take into account when drafting policy. These results generally demonstrate the efficacy of our approach using LLMs to generate scenarios, using those scenarios to simulate new scenarios in light of a policy, and then evaluating perceptions of impacts using human raters. We envision this process could be used by policy makers and researchers alike to elucidate some potential impacts and efficacy of a given piece of policy and to help understand the trade-offs in how the policy relates to different impacts. Before a policy is formally codified it could be used to workshop policies at a brainstorming stage and to stimulate potential avenues of how policies could turn out in practice. Being able to rank the impacts that may be impacted by a proposed policy could then be used to allocate scarce experimental resources to-

wards verifying the potential of a policy to mitigate an impact, such as through policy pilot programs.

Crucial to future work will be evaluating the approach we prototype here with policymakers in order to assess whether there are gaps that need to be bridged to make it useful in practice. Prior work underlines the importance of involving non-experts and laypeople in anticipatory assessments (Bonaccorsi, Apreda, and Fantoni 2020; Metcalf et al. 2021). However, critical judgment and expertise are necessary to confidently incorporate that information into larger processes of policy development. This includes getting (policy) expert opinions that validate the plausibility of the described policy effects in light of complex interactions between law, technology, and society.

In this study we only explored one part of one piece of legislation, but other policy proposals such as US President Biden's Executive Order on AI (Biden E.O. 14110 of Oct 30, 2023) or the Chinese Generative AI Regulation (Cybersecurity Administration of China No. 15, July 13, 2023) offer interesting potential points of comparison. We acknowledge that we are exploring the potential impact of EU legislation in a US context—however the EU AI Act is a specific implementation of some general ideas around labeling and transparency of generative AI systems and those general ideas also exist in draft form in some proposed legislation in US Congress. We chose to use the EU AI Act as it was a specific expression of the general ideas which are similarly applicable in the US context. Future work will need to extend this method to account for more complex policies as well as to consider how different policies may interact to produce impacts on different stakeholders in society and in different societies.

While we demonstrate this method with respect to negative impacts of generative AI in the news ecosystem, future work can be done to demonstrate generalizeability to other AI ethics harms, or harms even completely outside of AI. Through our experimentation, we discovered that utilizing our method alongside a human-made taxonomy of outlined harms provide for a more comprehensive understanding than allowing the LLM to come up with the harms itself. This comes against the cost of having a typology ready to go, but we found it was helpful for overall generation to ensure diverse impact themes emerged. Thankfully there are already rich typologies of harms of all sorts such as sociotechnical harms of algorithmic systems (Shelby et al. 2023), risks of harm from language models (Weidinger et al. 2021), or harms of generative audio models (Barnett 2023). Future work should also be conducted to assess the generalizeability of this method—for instance, when harms are extremely niche and not widely discussed in training datasets it is unclear whether LLMs may struggle to generate relevant scenarios for these types of impacts.

We further position the method established in this work with respect to the US NIST Risk Management Framework (NIST 2023)—a structure that organizations can use to facilitate risk management of design, development, and deployment of AI tools. The framework organizes risk management into three steps: mapping, measuring, and management. Creating the typologies, such as the one utilized in this paper constitute the mapping step, and then this method introduces a way to both measure the (perception of) severity and magnitude of these impacts to allow for stakeholders in power (e.g., legislators, government officials, AI researchers) to manage these harms. This method can assist with the prioritization of managing these harms by exploring the effectiveness of potential mitigation strategies which assists in deciding which risks to address first. Measuring severity, plausibility, magnitude, and specificity to vulnerable populations can also inform mitigation strategies beyond policy intervention. Our method is straightforward to adapt and can be extended and tested in new domains.

Specific Limitations of GPT-4 We want to further emphasize the limitations of this work that are derived from the use of GPT-4. Of the 50 impact types outlined by the Kieslich taxonomy, we were only able to generate 39 utilizing our prompting method (Appendix A.2.) When we framed the prompts more broadly in terms of the ten impact themes, the generated scenarios focused on a heavily skewed set of impact types (see Figure A5 in the Appendix A.3). Finally, we were entirely unable to generate some harms (e.g., discrimination), which we hypothesize could be due to the highly tuned nature of this model to prevent toxicity being present in outputs. There will be biases present in all LLMs—both due to tuning and training data. Future work needs to explore how those biases may be reflected in scenarios and in models' simulation of how a policy might impact a scenario.

Conclusion

In this work we develop and demonstrate the viability of a methodology to evaluate the potential mitigation effect of polices on given societal harms. We did so by using a large language model (specifically GPT-4) to generate scenarios in the near future both mitigated and non-mitigated by a given policy and then evaluate the perception of severity, plausibility, magnitude, and specificity to vulnerable populations. This methodology can support anticipatory governance approaches and risk management frameworks by lowering the cost for exploring policy options to mitigate harms from established taxonomies of impact. In addition it can help orient attention towards promising policy options where more rigorous or longitudinal evaluation is warranted and could furthermore be useful to help brainstorm the potential of different policies or mitigation strategies. Limitations of the method as well as a host of future work which can expand on the approach are discussed.

Acknowledgments

This research is partially supported by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence.

References

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*.

Alasadi, E. A.; and Baiz, C. R. 2023. Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education*, 100(8): 2965–2971.

Amer, M.; Daim, T. U.; and Jetter, A. 2013. A review of scenario planning. *Futures : the journal of policy, planning and futures studies*, 46: 23–40. ISBN: 0016-3287.

Amos-Binks, A.; Dannenhauer, D.; and Gilpin, L. H. 2023. The anticipatory paradigm. *AI Magazine*, 44(2): 133–143.

Anthropic. 2024. Claude 3 haiku: our fastest model yet.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The Moral Machine experiment. *Nature*, 563(7729): 59–64.

Barnett, J. 2023. The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 146–161.

Barnett, J.; and Diakopoulos, N. 2022. Crowdsourcing impacts: exploring the utility of crowds for anticipating societal impacts of algorithmic decision making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 56–67.

Benjamin, R. 2019. *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons.

Biden, J. E.O. 14110 of Oct 30, 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Bird, C.; Ungless, E. L.; and Kasirzadeh, A. 2023. Typology of Risks of Generative Text-to-Image Models. ArXiv:2307.05543 [cs].

Bonaccorsi, A.; Apreda, R.; and Fantoni, G. 2020. Expert biases in technology foresight. Why they are a problem and how to mitigate them. *Technological forecasting & social change*, 151. ISBN: 0040-1625.

Börjeson, L.; Höjer, M.; Dreborg, K.-H.; Ekvall, T.; and Finnveden, G. 2006. Scenario types and techniques: towards a user's guide. *Futures*, 38(7): 723–739.

Brey, P. 2017. Ethics of emerging technology. *The ethics of technology: Methods and approaches*, 175–191.

Burnam-Fink, M. 2015. Creating narrative scenarios: Science fiction prototyping at Emerge. *Futures*, 70: 48–55.

Buçinca, Z.; Pham, C. M.; Jakesch, M.; Ribeiro, M. T.; Olteanu, A.; and Amershi, S. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. ArXiv:2306.03280 [cs].

Börjeson, L.; Höjer, M.; Dreborg, K.-H.; Ekvall, T.; and Finnveden, G. 2006. Scenario types and techniques: Towards a user's guide. *Futures*, 38(7): 723–739. ISBN: 0016-3287.

Cai, H.; Liu, S.; and Song, R. 2023. Is Knowledge All Large Language Models Needed for Causal Reasoning? *arXiv preprint arXiv:2401.00139*.

Chanda, S. S.; and Banerjee, D. N. 2022. Omission and commission errors underlying AI failures. *AI & society*, 1–24.

Chen, Y.; Yang, X.-H.; Wei, Z.; Heidari, A. A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; and Guan, Q. 2022. Generative adversarial networks in medical image augmentation: a review. *Computers in Biology and Medicine*, 144: 105382.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Commission, E. 2024. 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, Pub. L. No. COM(2021) 206 final.

Cybersecurity Administration of China. No. 15, July 13, 2023. Interim Measures for the Management of Generative Artificial Intelligence Services.

Das, R.; Wong, Y. N.; Jones, R.; and Jackson, P. J. 2024. How do we speak about algorithms and algorithmic media futures? Using vignettes and scenarios in a citizen council on data-driven media personalisation. *New Media & Society*, 14614448241232589.

Diakopoulos, N.; and Johnson, D. 2021. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New media & society*, 23(7): 2072–2098.

Fuerth, L. 2011. Operationalizing Anticipatory Governance. *PRISM*, 2(4): 31–46. Publisher: Institute for National Strategic Security, National Defense University.

Glette-Iversen, I.; Aven, T.; and Flage, R. 2022. The concept of plausibility in a risk analysis context: Review and clarifications of defining ideas and interpretations. *Safety science*, 147: 105635.

Guston. 2014. Understanding 'anticipatory governance'. *Social Studies of Science*, 44(2): 218–242.

Hagerty, A.; and Rubinov, I. 2019. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892*.

Hoffmann, M.; and Frase, H. 2023. Adding Structure to AI Harm. Technical report, Center for Security and Emerging Technology.

Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Zhiheng, L.; Blin, K.; Adauto, F. G.; Kleiman-Weiner, M.; Sachan, M.; et al. 2023. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kıcıman, E.; Ness, R.; Sharma, A.; and Tan, C. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Kieslich, K.; Diakopoulos, N.; and Helberger, N. 2024. Anticipating impacts: using large-scale scenario-writing to explore diverse implications of generative AI in the news environment.

Kieslich, K.; Helberger, N.; and Diakopoulos, N. 2024. My Future with My Chatbot: A Scenario-Driven, User-Centric Approach to Anticipating AI Impacts.

Kıcıman, E.; Ness, R.; Sharma, A.; and Tan, C. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. ArXiv:2305.00050 [cs, stat].

Li, B. Z.; Nye, M.; and Andreas, J. 2021. Implicit representations of meaning in neural language models. *arXiv preprint arXiv:2106.00737*.

Long, S.; Piché, A.; Zantedeschi, V.; Schuster, T.; and Drouin, A. 2023a. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*.

Long, S.; Schuster, T.; Piché, A.; de Montreal, U.; Research, S.; et al. 2023b. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*.

Martin, L.; Whitehouse, N.; Yiu, S.; Catterson, L.; and Perera, R. 2024. Better Call GPT, Comparing Large Language Models Against Lawyers. *arXiv preprint arXiv:2401.16212*.

Martínez, E. 2024. Re-Evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, 1–24.

Metcalf, J.; Moss, E.; Watkins, E. A.; Singh, R.; and Elish, M. C. 2021. Algorithmic impact assessments and accountability: the co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Meßmer, A.-K.; and Degeling, M. 2023. Auditing Recommender Systems. Putting the DSA into practice wit a risk-scenario-based approach. Technical report, Stiftung Neue Verantwortung.

Moss, E.; Watkins, E.; Singh, R.; Elish, M. C.; and Metcalf, J. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. *SSRN Electronic Journal*.

Nanayakkara, P.; Diakopoulos, N.; and Hullman, J. 2020. Anticipatory ethics and the role of uncertainty. *arXiv preprint arXiv:2011.13170*.

Nikolova, B. 2014. The rise and promise of participatory foresight. *European Journal of Futures Research*, 2(1). ISBN: 2195-4194.

NIST. 2023. NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

Sarewitz, D. 2011. Anticipatory governance of emerging technologies. *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*, 95–105.

Selbst, A. D. 2021. An Institutional View of Algorithmnic Impact. *Harvard Journal of Law & Technology*, 35(1). Publisher: HeinOnline.

Selin, C. 2006. Trust and the illusive force of scenarios. *Futures : the journal of policy, planning and futures studies*, 38(1): 1–14. ISBN: 0016-3287.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. ArXiv:2210.05791 [cs].

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Daumé III, H.; Dodge, J.; Evans, E.; Hooker, S.; Jernite, Y.; Luccioni, A. S.; Lusoli, A.; Mitchell, M.; Newman, J.; Png, M.-T.; Strait, A.; and Vassilev, A. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. ArXiv:2306.05949 [cs].

Stahl, B. C.; Antoniou, J.; Bhalla, N.; Brooks, L.; Jansen, P.; Lindqvist, B.; Kirichenko, A.; Marchal, S.; Rodrigues, R.; Santiago, N.; Warso, Z.; and Wright, D. 2023. A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tu, R.; Ma, C.; and Zhang, C. 2023. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*.

Urueña, S. 2019. Understanding "plausibility": A relational approach to the anticipatory heuristics of future scenarios. *Futures*, 111: 15 – 25.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

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 *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Yao, B.; Jiang, M.; Yang, D.; and Hu, J. 2023. Empowering LLM-based machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.