A Demonstration of BLIP: A System to Explore Undesirable Consequences of Digital Technologies

ROCK YUREN PANG, SEBASTIN SANTY, RENÉ JUST, KATHARINA REINECKE, Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA

Undesirable consequences of digital technologies are often difficult to foresee at the time of their design and development, but having access to examples can help. This demonstration paper presents Blip, a research prototype that provides a catalog of undesirable consequences by extracting, summarizing, and categorizing undesirable consequences from online articles and academic papers using large language models (LLMs). Blip provides a web-based interactive user interface that allows users to explore undesirable consequences pertaining to different life aspects (e.g., “economy” or “politics”) and bookmark those found to be relevant.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**.

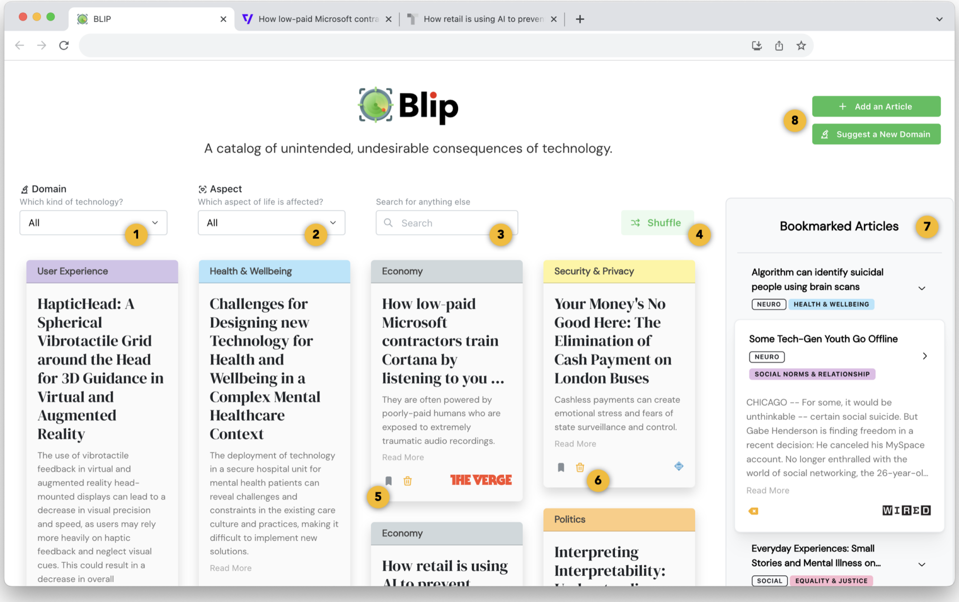
Additional Key Words and Phrases: Ethics, Unintended Consequences, Large Language Models (LLMs), Systems

ACM Reference Format:
Rock Yuren Pang, Sebastin Santy, René Just, Katharina Reinecke. 2024. A Demonstration of BLIP: A System to Explore Undesirable Consequences of Digital Technologies. In *29th International Conference on Intelligent User Interfaces - Companion (IUI Companion ’24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640544.3645237

1 INTRODUCTION

The advance of digital technologies have brought many benefits but also a range of undesirable consequences from hate speech in voice assistant systems [18, 19] to physical harms when users interact with virtual reality [16]. Researchers indicate that many of these negative effects on individuals and society could have been avoided if technology developers and researchers were aware of similar issues and had cautiously evaluated potential undesirable consequences beforehand [2, 9]. In response, prior work has suggested that a catalog of known cases might support considering the potential undesirable consequences for technology innovations [3].

This paper demonstrates Blip, a research prototype that offers a self-updating catalog of undesirable consequences of digital technologies. Specifically, Blip uses natural language processing (NLP) techniques to automatically extract real-world undesirable consequences of technology from online technology magazines and academic papers, summarizes and categorizes them, and presents them through an interactive, web-based interface. Blip’s goal is to enable users (such as CS researchers and practitioners, but also policymakers and the interested public) to efficiently explore the societal impacts of technology discussed in online articles, including how these affect different aspects of life and society, such as health, equality, or politics. Blip can be accessed at https://blip.labinthewild.org/. An evaluation of Blip showing the usefulness of its catalog of undesirable consequences for gaining awareness of, and brainstorming about, potential societal effects is described in [13].



HapticHead: A Spherical Vibrotactile Grid around the Head for 3D Guidance in Virtual and Augmented Reality

The use of vibrotactile feedback in virtual and augmented reality head-mounted displays can lead to a decrease in visual precision and speed, as users may rely more heavily on haptic feedback and neglect visual cues. This could result in a decrease in overall

Challenges for Designing new Technology for Health and Wellbeing in a Complex Mental Healthcare Context

The deployment of technology in a secure hospital unit for mental health patients can reveal challenges and constraints in the existing care culture and practices, making it difficult to implement new solutions.

How low-paid Microsoft contractors train Cortana by listening to you ...

They are often powered by poorly-paid humans who are exposed to extremely traumatic audio recordings.

How retail is using AI to prevent

Your Money's No Good Here: The Elimination of Cash Payment on London Buses

Cashless payments can create emotional stress and fears of state surveillance and control.

Interpreting Interpretability:

Algorithm can identify suicidal people using brain scans

Some Tech-Gen Youth Go Offline

CHICAGO -- For some, it would be unthinkable -- certain social suicide. But Gabe Henderson is finding freedom in a recent decision: He canceled his MySpace account. No longer enthralled with the world of social networking, the 26-year-ol...

Everyday Experiences: Small Stories and Mental Illness on...

Fig. 1. Blip’s main user interface.

2 THE USER EXPERIENCE

We illustrate how users can interact with Blip’s user interface to explore undesirable consequences. At a high level, the main interface (Figure 1) allows users to browse diverse examples of undesirable consequences of different technologies, understand and access the source articles, filter and search undesirable consequences, and bookmark articles, e.g. if they wanted to read an article later or create a subset of undesirable consequences.

The Blip interface facilitates user exploration of various technological impacts by presenting summaries of negative consequences in a card-based, scrollable gallery. Each card includes a color-coded header according to a distinct life aspect that it affects (e.g., “user experience” and “health & well-being”). The card also includes a header, followed by a summary of the undesirable consequences. Clicking on the article title opens a new browser tab that shows the original article. Users can bookmark or delete an article to and from the history sidebar ⑦. By default, Blip shows all cards in random order, but users can filter the cards by technology domain ① and/or by the aspect of life ②. In ③, users can search more relevant articles within Blip, such as “large language models” or “augmented reality.” A shuffle button at ④ allows users to mix cards randomly in the interface to enable serendipitous discovery of new ideas. Users can save it ⑤ to their bookmark field ⑦. When users think that they have already known a consequence in an article, they could remove that from their view ⑥. Users can look at their collection of articles at ⑦ to gain awareness of the consequences discussed online. Users can also import an article via an article URL in ⑧ as described below. In the future, we anticipate providing users with the option to share their own experiences with specific undesirable consequences that they find listed in Blip, or to add new ones.

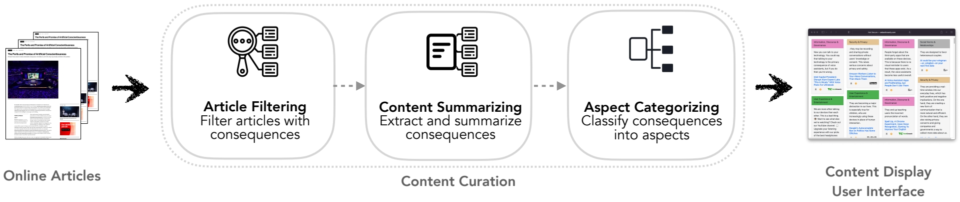


Fig. 2. An overview of Blip's content curation pipeline.

3 CONTENT CURATION

In Figure 2, we demonstrate Blip's content curation pipeline. Blip automatically filters relevant articles describing undesirable consequences of given input articles in a technology domain, extracts and summarizes these consequences, categorizes them into different aspects of life and society that they affect, and finally displays them in the interface in Section 2. For the steps in Figure 2, we employed model finetuning and prompting language models. Concretely, when prompting models, we used the gpt-3.5-turbo-instruct model, a pre-trained language model that can solve diverse NLP tasks with natural language instructions. We employed the 'zero-shot' methods in our case [17] since annotating articles (for supervised approaches) in all the steps is expensive because of the article length and relatively infrequent descriptions of undesirable consequences within them. The current version includes sources from the MIT Tech Review, WIRED, TechCrunch, and Futurism, as well as the papers from the CHI (2003-2023) and FAccT (2018-2023) proceedings.

3.1 Article Filtering

Given a large volume of input articles, filtering relevant articles that contain undesirable consequences is our first step. Blip filters the articles based on two types of information: the title and the content. This hybrid method aims to optimize efficiency and reduce the need for extensive OpenAI API queries.

First, the titles can often serve as a preliminary indicator of an article's relevance. For instance, the title "Social media is polluting society. Moderation alone won't fix the problem" is likely to discuss the social impact of social media. In contrast, titles that announce product launches (e.g., Apple Vision Pro) or staffing changes (e.g., "budget cuts and layoffs") are less likely to mention the technology's social repercussions (e.g., "Improbable teams with Google, opens SpatialOS alpha for virtual world development" [8]). To achieve the title-based filtering, Blip employs a RoBERTA-based [7] supervised binary classifier that outputs whether an article is relevant. To develop this title classifier, we annotated a dataset of 1,500 random online article titles to determine whether they are likely to contain an undesirable consequence or not.

Second, Blip filters the article content using the prompting approach of GPT-3.5 [1]. More precisely, Blip uses the following prompt: "Does the article above discuss unintended or undesirable consequences on society of <domain>? Answer Yes or No." See the Supplementary Materials for an example. To evaluate the filtering performance, Table 1 compared our choice of fine-tuned RoBERTa model for the title classifier and GPT-3.5 with only the RoBERTa model and a title classifier fine-tuned on a Natural Language Inference (NLI) task.

3.2 Content Summarizing

Blip automatically summarizes undesirable consequences in the filtered articles. Reading the entire article is time-consuming. We found it difficult to skim through paragraphs and identify phrases describing the consequences without getting sidetracked by details. Hence, compiling a list of consequences in articles was largely impractical because irrelevant details in the articles distracted from higher-level issues. Instead, a shorter summary of the discussed

Table 1. Performance of Article Filtering

Relevance Classifier	F1	Acc.
Title [always irrelevant]	0.00	0.61
NLI (zero-shot)	0.43	0.62
RoBERTa (supervised)	0.87	0.87
gpt-3.5-turbo (zero-shot, ours)	**0.90**	**0.89**

consequence helps the user grasp the overall issue. Blip employs GPT-3.5 for abstractive summarization, which paraphrases the undesirable consequences discussed in the article and generates relatively short summaries. We leveraged LLMs given that prior work found LMM summaries to be on par with human written summaries [20]. The prompt for this task is: "To summarize in a short paragraph, the main undesirable consequence of <domain> being discussed here is". See the Supplementary Materials for an example.

3.3 Aspect Categorizing

Blip assigns each undesirable consequence summary to one of 10 aspects of life, from health & well-being to politics. This categorization narrows down the set of summarized undesirable consequences and can emphasize the variety of impacts technology can have on society. To develop the list of aspects, we built on the 21 aspects of life relevant to societal implications from the Artefact Group's Tarot Cards of Tech project [4]. Assigning 150 randomly chosen articles discussing undesirable consequences from our dataset to these 21 aspects of life, we iteratively merged and renamed the aspects to fit our data better. The resulting 10 aspects of life broadly represent various categories that undesirable consequences commonly fall into and are used in Blip to support users in learning and brainstorming. We incorporated the list in Blip such that it can be extended with additional aspects or replaced with a new list. Blip uses the prompting approach of GPT-3.5 for aspect categorization. The prompt we use for this task is: "Which aspect of life does the following consequence affect?" See the Supplementary Materials for an example.

3.4 Implementation Details

Blip includes a frontend interface implemented in the React JavaScript library and a server using the FastAPI Python framework. The server uses Selenium [5] and Beautifulsoup [14] to extract article URLs based on input keywords and the newspaper3k API [11] to obtain the article content. We used a combination of the sentence-transformers model in the huggingface library and the FAISS library to enable the quick search functions for similar articles to a search keyword [6]. In our main system architecture, we initially used GPT-3 and updated the content with the GPT-3.5 model [10]. The language models that Blip uses can be exchanged as more powerful versions come out. We also added an option to run the pipeline using open-source language models, Llama2 [15]. Currently, Blip includes 10 domains (i.e., Social Media, Voice Assistants, Augmented/Virtual Reality, Computer Vision, Robotics, Mobile, AI Decision-Making, Neuroscience, Computational Biology, and Ubiquitous Computing). These articles were extracted from MIT Tech Review, WIRED, TechCrunch, and Futurism, as well as the papers from the CHI (2003-2023) and FAccT (2018-2023) proceedings. Blip automatically extracts from the online magazines and updates the content weekly.

4 CONCLUSION AND FUTURE WORK

This demonstration presented Blip, a research prototype that aims to facilitate the exploration of undesirable consequences of digital technologies. Blip illustrates a unique approach for extracting undesirable consequences of technology from online text data at scale. Blip, including its content curation pipeline and the interactive interface, is fully open-source and available at https://blip.labinthewild.org/. In this work, we recognize that LLMs can introduce serious undesirable consequences, such as model hallucination, bias in the training data, and limited understanding of emerging fields. To address these concerns, our work extracts relevant information directly from trusted sources (i.e., online articles and papers) and provides access to the original content, instead of directly prompting LLMs to generate the information. While an evaluation of Blip has shown promising results in supporting CS researchers in learning and brainstorming about undesirable consequences [13], future work needs to evaluate how Blip could be integrated into the technology research, design, and development process. Future work is also required to explore whether Blip could, over time, help to mitigate undesirable consequences [12]. In addition, our initial evaluations presented in [13] suggest that browsing the catalog of undesirable consequences makes users want to add their thoughts and personal experiences. As such, Blip may provide exciting new opportunities for citizen scientists to document their personal experiences, which we plan to test in future work. We hope that Blip will spark new conversations in the IUI community about the design of tools that facilitate the exploration of undesirable consequences of digital technologies and that, ultimately, support researchers, designers, and developers in mitigating the negative effects of technology on society.

ACKNOWLEDGMENTS

We thank our participants and the anonymous reviewers for their valuable feedback and suggestions. This work was funded by the National Science Foundation under award IIS-2006104 and SMA-2315937. Any opinions, findings, conclusions, or recommendations expressed in our work are those of the authors and do not necessarily reflect those of the supporter.

REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] Amy Bruckman. 2020. ’Have You Thought About…’: Talking about Ethical Implications of Research. *Commun. ACM* 63, 9 (aug 2020), 38–40. https://doi.org/10.1145/3377405

[3] Kimberly Do, Rock Yuren Pang, Jiachen Jiang, and Katharina Reinecke. 2023. “That’s Important, but...”: How Computer Science Researchers Anticipate Unintended Consequences of Their Research Innovations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI ’23)*. Association for Computing Machinery, New York, NY, USA, Article 602, 16 pages. https://doi.org/10.1145/3544548.3581347

[4] Artefact Group. 2017. The Tarot Cards of Tech. last accessed October 18, 2023.

[5] Jason Huggins. 2018. *Selenium with Python.* Retrieved 2021-04-06 from https://selenium-python.readthedocs.io/

[6] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[8] Ingrid Lunden. 2016. Improbable teams with Google, opens Spatialos Alpha for virtual world development. https://techcrunch.com/2016/12/13/improbable-teams-with-google-opens-spatialos-alpha-for-virtual-world-development/

[9] Jeanna Matthews. 2022. Embracing Critical Voices. *Commun. ACM* 65, 7 (Jun 2022), 7. https://doi.org/10.1145/3535268

[10] OpenAI. [n. d.]. Models. https://platform.openai.com/docs/models/gpt-3

[11] Lucas Ou-Yang. 2013. *Newspaper3k: Article scraping & curation.* Retrieved 2022-04-06 from https://newspaper.readthedocs.io/en/latest/

[12] Rock Yuren Pang, Dan Grossman, Tadayoshi Kohno, and Katharina Reinecke. 2023. The Case for Anticipating Undesirable Consequences of Computing Innovations Early, Often, and Across Computer Science. arXiv:2309.04456 [cs.CY]

[13] Rock Yuren Pang, Sebastin Santy, René Just, and Katharina Reinecke. 2024. BLIP: Facilitating the Exploration of Undesirable Consequences of Digital Technologies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, Hawai’i, USA) *(CHI ’24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642054

[14] Leonard Richardson. 2020. *Beautiful Soup Documentation*. Retrieved 2021-04-06 from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[15] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[16] Wen-Jie Tseng, Elise Bonnail, Mark McGill, Mohamed Khamis, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2022. The Dark Side of Perceptual Manipulations in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 612, 15 pages. https://doi.org/10.1145/3491102.3517728

[17] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models Are Zero-Shot Learners. *ArXiv* abs/2109.01652 (2021).

[18] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[19] Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications. *The ORBIT Journal* 1, 2 (2017), 1–12.

[20] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. arXiv:2301.13848 [cs.CL]