Blip: Facilitating the Exploration of Undesirable Consequences of Digital Technologies

Rock Yuren Pang
ypang2@cs.washington.edu
Paul G. Allen School of Computer Science,
University of Washington
Seattle, Washington, USA

Sebastin Santy
ssanty@cs.washington.edu
Paul G. Allen School of Computer Science,
University of Washington
Seattle, Washington, USA

René Just
rjust@cs.washington.edu
Paul G. Allen School of Computer Science,
University of Washington
Seattle, Washington, USA

Katharina Reinecke
reinecke@cs.washington.edu
Paul G. Allen School of Computer Science,
University of Washington
Seattle, Washington, USA

ABSTRACT

Digital technologies have positively transformed society, but they have also led to undesirable consequences not anticipated at the time of design or development. We posit that insights into past undesirable consequences can help researchers and practitioners gain awareness and anticipate potential adverse effects. To test this assumption, we introduce Blip, a system that extracts real-world undesirable consequences of technology from online articles, summarizes and categorizes them, and presents them in an interactive, web-based interface. In two user studies with 15 researchers in various computer science disciplines, we found that Blip substantially increased the number and diversity of undesirable consequences they could list in comparison to relying on prior knowledge or searching online. Moreover, Blip helped them identify undesirable consequences relevant to their ongoing projects, made them aware of undesirable consequences they "had never considered," and inspired them to reflect on their own experiences with technology.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

undesirable consequences, computer ethics, societal impacts, NLP

ACM Reference Format:
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 CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642054

1 INTRODUCTION

With the rise of digital technologies in our lives, society has not only experienced their benefits but also increasingly their undesirable consequences. Research and news headlines describe seemingly unavoidable side effects of digital technologies, from Instagram's adverse effects on adolescent girls' body images [52] to Microsoft's chatbot Tay using racist language [112]. Technological progress is commonly seen as a moral commitment that is "legitimized no matter how dangerous" [83, p.325]. While undesirable consequences are sometimes described as accidental and minor *"blips"*, researchers, journalists, and policymakers have suggested that many cases could have been avoided if technology developers were aware of similar issues and had taken cautious evaluation beforehand [14, 61]. However, anticipating the various outcomes of technology is difficult [70]. In fact, recent work has found that computer science (CS) researchers at the forefront of developing new technologies are eager to proactively consider undesirable consequences of their innovations, but lack well-formulated processes and tools to do so effectively [31]. They reported that not having resources that provide a comprehensive understanding of "common problems" reduced their ability to anticipate undesirable consequences [31, p.7]. Could insights into past undesirable consequences of technology help them gain awareness of potential future consequences?

We study this question by collecting a catalog of "common problems," allowing CS researchers to explore past undesirable consequences as reported in technology magazines and research papers. Learning from prior incidents has proven to be useful in several settings, ranging from exploratory forecasting of technological advances [60], improving software by studying a collection of previous defects [49], to training future pilots using the aviation accidents database [55]. Incorporating known and real-world case studies of ethical dilemmas into an undergraduate Human-Centered Computing class has also been shown to amplify students' engagement in ethical thinking [98]. What remains unknown is whether providing CS researchers with examples can have similar advantages, increasing their awareness of the various societal impacts of technology and supporting them in considering the potential consequences of their own projects. A challenge in the domain of undesirable consequences is the lack of such resource across diverse CS subfields, and the unclear impact on CS researchers, who often lack the time for such in-depth consideration for diverse consequences [31, 99].

Hence, the goal of this paper is to explore whether providing CS researchers with a catalog of "common problems" would improve their awareness of undesirable consequences. Our secondary goals are to find out how we can feasibly collect a self-updating catalog

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642054

given that this information is currently scattered and the technology landscape is fast-moving, and how a system providing this service would be perceived and used by CS researchers. We tackle these questions by designing, developing, and evaluating Blip, a prototype system that collects and showcases *a catalog of undesirable consequences of digital technologies*. Blip (1) automatically extracts real-world undesirable consequences of technology from any given online article using natural language processing (NLP) techniques, (2) summarizes and categorizes them based on the aspect of life that they affect (such as health, equality, or politics), and (3) presents them in an interactive, web-based interface (see Figure 1). Users can use Blip to view, sort, and save the currently 5.7k summaries of undesirable consequences, or extract undesirable consequences from additional articles.

While considering undesirable consequences is not yet a common practice among researchers, we designed Blip to facilitate this process in the future. We tested our assumption about Blip's usefulness in two user studies. In the first study with nine CS researchers, we assessed Blip's overall usefulness to consider undesirable consequences in their broader field (e.g., social media) compared to two alternative approaches—relying on their prior knowledge and searching for undesirable consequences online—and conducted in-depth interviews to further understand users' perceptions of Blip and potential use cases. Our results show that Blip enabled participants to add an average of 7.00 more undesirable consequences beyond those they could list when relying on prior knowledge and searching online. Participants perceived Blip as improving their ability to "think outside the box", made them aware of consequences that they "had never considered," and was an essential way to collect undesirable consequences "because you can't just read a bunch of disconnected articles about this [and make sense of it]."

In our second study with six CS researchers, we followed up on these results, evaluating whether Blip is useful and actionable in the context of specific projects that participants work on across CS subdisciplines. All participants could find several undesirable consequences relevant to their specific projects in less than 15 minutes, on average. Some of these were immediately actionable. Overall, this paper contributes:

(1) Empirical evidence that a catalog of undesirable consequences supports CS researchers in considering more, and more diverse, undesirable consequences than if they rely on their prior knowledge or an online search (Study 1) and that it enables them to uncover potential adverse effects of their own projects (Study 2).
(2) An open-source, web-based system, Blip[1], that collects, summarizes, and categories undesirable consequences. To develop Blip, we designed an information distillation pipeline that leverages NLP techniques to efficiently establish a self-updating catalog of undesirable consequences.

2 RELATED WORK

We use the term "undesirable consequences" to refer to negative consequences of digital technology that affect society [27]. Oftentimes, undesirable consequences are unanticipated or even unintended [65]. We chose to work with "undesirable consequences" over the more prevalent "unintended consequences" of technology to emphasize that our primary concern is with exploring the adverse effects of technology. Research in HCI and Science and Technology Studies (STS) has contributed a large body of work on observed negative effects of digital technology domains and products, including mobile phones [69, 87], the sharing economy [30], machine learning [18, 19], and social media [28, 100]. Researchers have also described various aspects of our lives that may be adversely affected by technology, such as its impacts on the environment [12], health [5, 43], or privacy [3]. Moreover, researchers have increasingly investigated and brought to our attention differential effects of digital technology on certain population groups, such as on different gender [32] and racial groups [17, 93], low-income and underserved communities [30], or people in other countries [79, 86, 92, 106].

Discussions and interventions for addressing undesirable consequences in research. With the increasing awareness of the potential adverse effects of technological innovations, the research community has started to engage in several efforts to prevent such incidents. For example, researchers have developed guidelines for ethical research and development [4, 104], started dedicated conferences, such as FAccT, AIES, EAAMO, SIGCAS, and dedicated tracks (e.g., Critical Computing@CHI), led workshops [102] and ethical committees [20, 35], as well as called for changes in institutional structures [10, 80], critical education [53], and in how we address undesirable consequences of digital technologies [14, 61]. A key concrete step was the inclusion of broad ethics or impact statements in major conferences, such as IUI [1], NeurIPS [11], ACL [101]. Nanayakkara et al. [71] found that such statements diversified thinking about how ML research could potentially impact society, though they tended to focus on positive impact [6]. Additionally, there are calls for researchers within different computing communities to accurately report the design considerations of their datasets [9, 38], models [67, 89], and tasks [56, 68] as well as evaluate and de-bias their products [4, 8, 104, 111].

Methods for forecasting and anticipating undesirable consequences. Researchers have designed tools to help identify and contemplate social values of different stakeholders, such as the Envisioning Cards [36], Tarot Cards of Tech [41], and Value Cards [94] (see also [22] for a detailed overview). The value-sensitive design approach has also contributed broad guidelines for researchers seeking to account for human values in a principled and systematic manner throughout the design process [37, 118]. The Future Ripples method [34], inspired by the Futures Wheel foresight method in education [39], allows collaborative brainstorming on the impact of innovation through workshop activities.

However, some of these approaches and methods have been challenged for not sufficiently supporting practitioners and the reality of the product-development process [40]. Using these methods requires prior knowledge on the topic at hand, which may not always be the case for novice users. They also require developers to deliberate in a team, sometimes with external experts, simultaneously and collectively where envisioned consequences can vary depending on the team's diversity and backgrounds. In fact, an interview with 20 CS researchers found that none of the participants are actively using these tools in practice [31].

[1] https://blip.labinthewild.org/

An alternative approach for anticipating undesirable effects of technology is learning from past incidents [64, 114]. While perfectly predicting the future may be impossible, researchers have developed various methods to estimate what may happen from such past experiences. For example, the Delphi forecasting method [95, 109] has been used in a wide variety of domains such as predicting air travel [33] and designing educational technology [76]. Another forecasting method is the case study method, which collects people's thoughts on and experiences with past technology developments in an organization [21]. One issue for the widespread use of these methods is that they usually rely on experts to collect and interpret historical data, making it difficult to scale and frequently use them.

In another attempt to anticipate undesirable consequences, prior work has developed a forum to collect news articles about technologies [73] and an AI incident database specifically for the effects of AI technology on society [64]. One of the motivations for this database was that "the artificial intelligence system community has no formal systems whereby practitioners can discover and learn from the mistakes of the past" [64, p.1]. However, the AI incident database necessitates the crowd to manually browse and enter incidences, using a leaderboard to incentivize contributions from volunteers. To the best of our knowledge, no previous system exists that automatically and systematically catalogs, summarizes, and categorizes undesirable consequences for a variety of technology domains. None of these approaches have been formally evaluated to show their usefulness for anticipating undesirable consequences.

In short, many prior tools prompt users to reflect on high-level ethical questions, which requires users to have prior knowledge without easy access to updated real-world examples. This paper explores the value of providing researchers with concrete examples that can ground their ethical considerations in practice.

Supporting ideation of potential undesirable consequences. Blip was also inspired by work on creativity support tools, which showed that a collection of diverse examples can support ideation [81, 97]. For example, sampling diverse inspirational examples (and providing a visual overview of the ideas) has been found to improve people's brainstorming activity [96]. Similar work on cognition and creativity support confirmed that examples inspire and unveil new and diverse ideas [29, 51, 74, 75, 116]. To organize these examples, prior work leveraged categories of certain topics and characteristics, which are essential to human cognition [90, 108]. For example, IdeaRelate facilitates the exploration of COVID-related examples by tagging them into different topics, helping users to include more perspectives in their own idea generation [113]. Recent work attempted to incorporate language models to help ideate potential harms [16, 82]. In particular, AngleKindling used few-shot LLM prompts to find potential controversies and negative outcomes from press releases to help journalists generate story ideas [84]. However, zero-shot and few-shot approaches to generate consequences had resulted in rather generic results [84]. In this work, we aid the inherently creative process of reflecting on past and possible future adverse effects by providing a catalog of undesirable consequences, supplemented with information on the diverse aspects of life that they have affected. Instead of relying on ideas generated entirely from language models, we extract relevant information directly from a wide range of online articles, provide access to the original content, and update our collection every week.

3 BLIP

Blip was developed to explore the value of providing researchers with a catalog of past undesirable consequences of technology. Showing this catalog aims to address gaps where developers often overlook potential adverse impacts [31]. Blip's open-source code is available at https://github.com/rrrrrrockpang/blip, and its interface is currently deployed at https://blip.labinthewild.org.

3.1 Design Choices and Rationale

We followed a user-centered design process in which we iteratively sought user feedback on several prototype interfaces before arriving at the present implementation. The design choices were also informed by prior literature as follows:

Everything in one place: Combining information across the Internet is difficult, which is why several systems address the need to collect information in one place (e.g., Pinterest, Fuse [54]). Blip is therefore designed as a web-based system that allows viewing and organizing undesirable consequences across various technology domains in one place.

Automatically collecting information: Prior work showed that developers desire a collection of past technology incidences, but that they lack the time to invest in collecting resources themselves [31]. Moreover, manual curation takes time and requires motivating users to contribute this data, which can be difficult and result in a limited number of undesirable consequences examples. To support the scale needed to achieve a fairly comprehensive and updated collection, we developed an approach for *automatically* retrieving undesirable consequences from a set of trusted online articles that regularly report on them. Blip currently retrieves articles from reputed outlets that often report on new technologies, such as MIT Technology Review, TechCrunch, The Verge, and WIRED. This list can be easily expanded.

Summarizing undesirable consequences: To help users effectively process online information, prior systems have summarized complex content in other contexts such as for reading papers [7], reviewing academic literature [50] and conversing online [115]. Similarly, we present users with a summary of any undesirable consequences in an article. Our decision was further informed by a design and feedback session with three CS researchers, in which we presented early mock-ups and discussed potential changes. Participants noted that the original articles were too long for a quick overview, though all wanted to retain the possibility to access them. Participants noted that seeing entire paragraphs prevented them from quickly understanding what the societal implication is. Instead, we decided to provide a summary specific to the undesirable consequences in an article.

Categorizing undesirable consequences: Prior work has examined the benefits of category structure in human cognition for sensemaking and creativity [108]. More recent systems have leveraged different "categories" to organize mass information [54] and generate creative ideas [74, 96, 113]. In addition, our decision was confirmed in the same preliminary study above, in which participants reported that presenting undesirable consequences without any organization was overwhelming and time-consuming. To address this issue, we designed Blip to categorize undesirable consequences into different aspects of life that they affect, such as politics and equality, which we adapted from the Tarot Cards of

Tech [41]. Blip visually signals these categories with different colors, which may enhance users' understanding of the diverse range of undesirable consequences that can occur in the real world.

Bookmarking articles: In early feedback on our prototype, we also repeatedly received the feedback that users wanted to return to a specific article or save it as a collection of undesirable consequences particularly relevant to their project. Blip therefore allows bookmarking articles in a sidebar using cookies.

3.2 User Interface Usage Scenario

In this section, 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 (1) browse through diverse examples of undesirable consequences for different technologies, (2) understand and access the source articles, (3) filter and search undesirable consequences, and (4) bookmark articles, e.g. if they wanted to read the article later or create a subset of undesirable consequences for later consideration.

Blip displays different undesirable consequences on cards in a scrollable interface (see Figure 1). As the user scrolls down, new cards appear automatically. Each card includes a header that displays the aspect of life it affects in a distinct color to promote visual organization. The card content includes the summarized undesirable consequence along with the article title and source, as well as two buttons that let users bookmark or delete an article from the view. Clicking on the article title opens a new browser tab that shows the original article. Bookmarked cards appear on a 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. They can also search for specific terms within the summary, such as "mental health" or "misinformation" 3. The shuffle button at the top allows users to shuffle cards in the collection view to encounter new ideas 4. Users can save a card 5 to their bookmark 7. When users think that they have already known a consequence in a card, they can remove that from their view 6. Users can review 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 in Section 3.5.

3.3 Content Curation Pipeline

As shown in Figure 2, Blip automatically filters relevant articles describing undesirable consequences of given input articles in a technology domain, (Section 3.3.1), extracts and summarizes these consequences (Section 3.3.2), categorizes them into different aspects of life and society that they affect (Section 3.3.3), and finally displays them in an interactive interface in Figure 1. To achieve these steps, Blip uses GPT-3.5 [13] due to its versatility and high-quality outputs. GPT-3.5 [13] is noted for its ability to classify with higher accuracy than supervised approaches with no or few training instances. We used the gpt-3.5-turbo model, a pre-trained language model that can solve NLP tasks with instructions. The model can be accessed via its OpenAI API [77]. To show the model how to perform a given task, it has to be given instructions along with examples. Such 'zero-shot' methods benefit our case [110] since annotating articles (for supervised approaches) is expensive because of the length of articles and relatively infrequent descriptions of undesirable consequences. Hereon, we use the terms INPUT to denote the input text, PROMPT to denote the natural language instruction, and OUTPUT to denote the output by the model.

3.3.1 Article Filtering. Given a large volume of input articles, filtering relevant articles that contain undesirable consequences is our first step. Blip performs filtering in two steps based on: (1) the title and (2) the content. This hybrid filtering method aims to include more relevant undesirable consequences from articles and reduces the cost of computation from requesting the OpenAI API.

Filtering by Title. First, Blip determines whether an article mentions undesirable consequences based on the title information. For example, the title "Social media is polluting society. Moderation alone won't fix the problem" is very likely to discuss such consequences. In contrast, titles that announce product launches, analyze products, or discuss corporate leadership rarely contain relevant consequences (e.g., "Improbable teams with Google, opens SpatialOS alpha for virtual world development" [58]).

To filter articles by title for those that mention undesirable consequences, Blip employs a RoBERTA-based [57] supervised binary classifier that outputs whether an article is relevant or not. To develop this title classifier, we annotated a dataset of 1,500 random online article titles for whether they are likely to contain an undesirable consequence or not. Two authors individually annotated all the article titles with a binary label "relevant" or "irrelevant." The initial inter-rater reliability was 92.17%. The two authors then discussed the inconsistent titles until agreement was achieved. When the two authors were unsure about the relevance during the deliberation phase, we included the titles and resorted to filtering by content to reduce ambiguity. Because obtaining diverse positive examples is more difficult than getting negative examples, we leveraged the AI incidents database [64] to find articles that discuss undesirable consequences. The title classifier that was fine-tuned on this dataset achieved F1=86.63% on a 4:1 train-test split. The performance of our title classifier is significantly higher than that of classifiers in comparable prior work (e.g., when detecting propaganda in news articles, where an F1 score of 60.98% was reported [26]).

Filtering by Content. Blip additionally filters the article content using the prompting approach [13]. More precisely, Blip uses the following PROMPT: "Does the article above discuss unintended or undesirable consequences on society of <domain>? Answer Yes or No." An example INPUT, PROMPT, and OUTPUT looks as follows:

An Example of Filtering by Content

The Nauseating Disappointment of Oculus Rift (MIT Tech Review) [66]

```
Admittedly, I was using a $599 Oculus Rift virtual-reality headset. It was a lot of fun, though I looked like a complete idiot sitting with a clunky black gadget on my face. I also got a more in-depth look at simulator sickness—feelings of nausea, dizziness, and eye strain that some people get when using VR—and what it means for the future of this technology ... [continued]
Does the article above discuss undesirable consequences of virtual reality on society? Answer Yes or No.
Yes
```

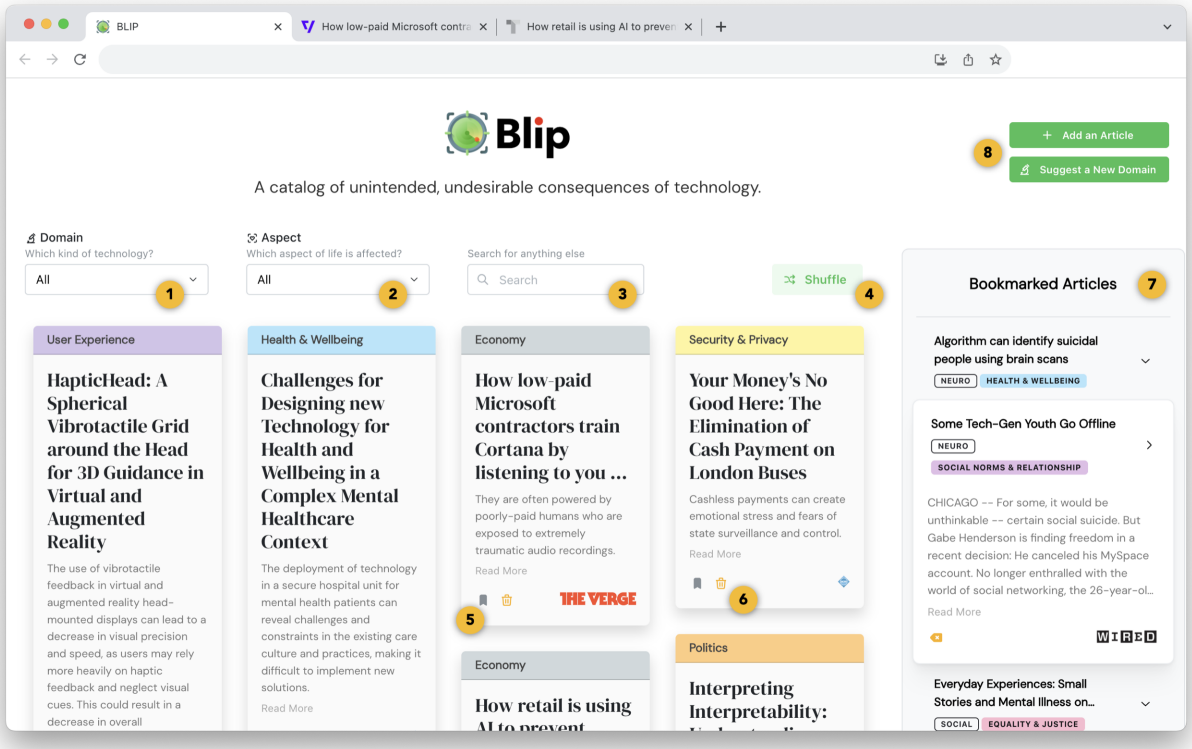

Figure 1: Blip's main user interface. Users can view summaries of undesirable consequences and filter them by technology domains, aspects of life they affect, or search keywords.

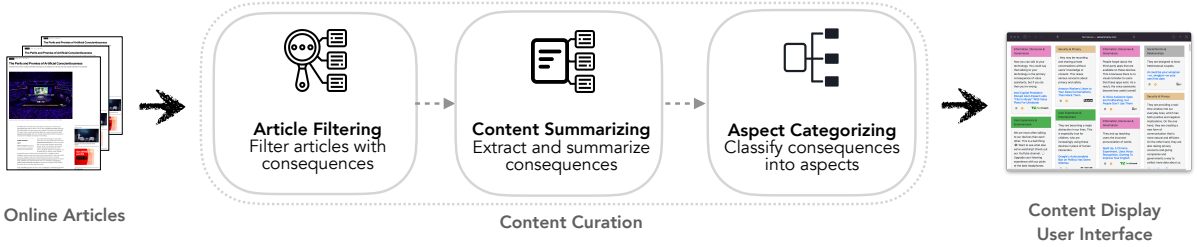

Figure 2: An overview of Blip's content curation pipeline. Given online article sources, Blip filters out those that discuss undesirable consequences, extracts and summarizes the consequence, and categorizes it into different aspects of life that it affects, such as the environment, equality, or politics. The undesirable consequences are then displayed in an interactive, web-based user interface.

3.3.2 Content Summarizing. Given the filtered set of articles, Blip automatically summarizes undesirable consequences. For example, consider the following paragraph from an article in the *MIT Technology Review* [91]:

> *[...] social media was just making it worse. The prejudice Lise experienced—colorism—has a long history, driven by European ideals of beauty that associate lighter skin with purity and wealth, darker tones with sin and poverty. [...] And today, thanks to the prevalence of selfies and face filters, digital colorism has spread. With Snapchat, Instagram, TikTok, and Facebook a part of billions of people's everyday lives, many of us find that people see far more pictures of us than ever before. [...]*

While providing much detail and nuance, reading the entire article is time-consuming. We found that articles in our dataset included an average of 713.2 words (min=103; max=5401; SD=501.6). In our experience, we found it challenging to skim and locate phrases within paragraphs that describe undesirable consequences without getting caught up in the specifics. Therefore, listing various undesirable consequences discussed in articles was impractical at best because irrelevant details in the articles distracted from higher-level issues. Instead, a shorter summary of the discussed consequence helps users 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 LLM

summaries to be on par with human written summaries [117]. In contrast to extractive summarization, which extracts and concatenates excerpts from an article into a summary, abstractive summarization can generate summaries of a specified length and include multiple consequences, making it a better fit for our purpose. The PROMPT for this task is: "To summarize in a short paragraph, the main undesirable consequence of <domain> being discussed here is". An example INPUT, PROMPT, and OUTPUT is as follows:

An Example of Content Summarizing

How digital beauty filters perpetuate colorism (MIT Technology Review) [91]

When Lise was a young teenager in Georgia, her classmates bullied her relentlessly. She had moved with her family from Haiti a few years earlier, and she didn't fit in with the other students. They teased her about her accent, claimed she "smelled weird," and criticized the food she ate. But most often they would attack her with remarks about her dark complexion ... As she struggled to cope with the abuse, Lise convinced herself that the darkness of her skin was to blame. And social media platforms and the visual culture of the internet suggested the same thing ... [continued]
To summarize in a short paragraph, the main undesirable consequence of social media being discussed here: It can lead to the reinforcement of colorism. Colorism is defined as prejudice against people with darker complexions, and it can have harmful effects on people's mental and physical health. Social media platforms like Instagram have filters that can lighten users' skin tone, which can perpetuate the idea that lighter skin is more desirable. In addition, recommendation algorithms on these platforms often favor content featuring people with lighter skin, which can reinforce users' biases and lead to the marginalization of people with darker complexions.

3.3.3 Aspect Categorizing. BLIP assigns each undesirable consequence summary to one of 10 aspects of life, from health & well-being to politics (see Supplementary Materials for a list of aspects and examples). This categorization allows for narrowing 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 list of 21 aspects of life that are relevant to societal implications from the Tarot Cards of Tech project [41]. Assigning 150 randomly chosen articles discussing undesirable consequences to these 21 aspects of life, we iteratively merged and renamed the aspects to fit our data (see Supplementary Materials for details). 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?" An example INPUT, PROMPT, and OUTPUT looks as follows:

An Example of Aspect Categorizing

AI voice actors sound more human than ever—and they're ready to hire (MIT Technology Review) [42]

List of possible aspects: Economy, Environment & Sustainability, Equality & Justice, Information & Discourse, Health & Well-being, Politics, Power, Security & Privacy, User Experience & Entertainment, Social Norms & Relationships
Which aspect of life does the following consequence affect?
Title: AI voice actors sound more human than ever—and they're ready to hire
Summary: People are losing their jobs. The technology is becoming so realistic that many people can't tell the difference.
Aspect: Economy

3.3.4 Implementation Details and Costs. BLIP includes a frontend interface implemented in the React JavaScript library and a server using the FastAPI Python framework. The server uses Selenium [46] and Beautifulsoup [88] to extract article URLs based on input keywords and the newspaper3k API [78] 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 [48]. In our main system architecture, we initially used the GPT-3 API and text-davinci-002 model [77], which was released on June 11, 2020. The cost of using the GPT-3 API was $0.06 for 750 words at the time of implementation. Since then, new variants of GPT models were made available, including GPT-3.5 and GPT-4, which were introduced after our first study. The language models that BLIP uses can be changed as more powerful versions come out. We also added an option to run the pipeline using open-source language models, Llama2. We re-ran our content curation pipeline on the three domains using GPT-3.5 on August 15, 2023.

3.4 Technical Evaluation

We evaluated the pipeline described above on three technology domains: social media (SM), virtual reality (VR), and voice assistants (VA). We chose these three domains as the initial content for BLIP because they represent diverse digital technologies that have been deployed and used for different amounts of time. Social media is a technology with widely-explored consequences on economic, political, and social spheres (e.g., polarization [107] and depression [25]). Voice assistants are comparatively new but are now an integral part of many people's lives, with consequences including privacy violations [72] and harmful content [47]. Virtual reality is still newer and has not yet become mainstream.

Retrieving Online Articles. We searched for articles in these domains using the keywords below from the sources in Table 1. The search led to a total of 42,405 articles, published between 1997-2023.

Article Filtering. Applying the title classifier to our dataset of online articles resulted in a total of 26,628 articles as shown in Table 1. Filtering by content retained 2.6k articles in our dataset

Table 1: Online sources for retrieving articles on the three technology domains in our technical evaluation: Social Media (SM), Voice Assistants (VA), and Virtual Reality (VR). The table shows the percentage and total number of relevant articles that contain consequences after each filtering step.

News Source	Retrieved Articles SM·VA·VR	Title Filter SM·VA·VR	Content Filter SM·VA·VR
MIT Tech Review 1997-2022	3433 1686·957·790	1957 (57%) 1082·563·312	519 (15%) 349·116·54
TechCrunch 2005-2022	3975 748·1502·1725	1330 (33%) 337·538·455	390 (10%) 155·187·48
The Verge 2011-2022	720 89·473·158	236 (33%) 61·160·15	175 (24%) 53·114·8
WIRED 2010-2022	34000 5345·17319·11516	22940 (67%) 3954·11560·7426	1489 (4%) 921·409·159
Total 1997-2022	42405 7968·20148·14289	26628 (63%) 5503·12855·8270	2616 (6%) 1498·840·278

that discuss undesirable consequences of SM, VA, and VR. This process filters out articles with ambiguous titles related to undesirable consequences. For example, an article titled "Advertisers Employ Social Media" was predicted as relevant by title filtering. However, the article discusses companies that use social media for advertisement with no clear undesirable consequences. The content-based classifier achieved an accuracy of 89.24% (F1=89.83%), which is a 3% increase compared to the title classifier.

- **Social Media**: `social media`[2]
- **Voice Assistants**: `voice assistant, chatbot, home assistant, AI assistant, speech recognition, voice recognition, smart assistant, personal assistant`
- **Virtual Reality**: `virtual reality, mixed reality, augmented reality, metaverse`

Content Summarizing. To evaluate the accuracy of the generated summaries in describing the undesirable consequences from the articles, we randomly chose 50 articles from the filtered set. One author read each article and graded the corresponding summary as either accurate or inaccurate. A second author then confirmed the decision. The summary was considered accurate if it truthfully translated the desired content (undesirable consequences in our case). Sources of inaccuracy included (1) introducing facts absent in the original text (also known as model hallucinations [62]), (2) failing to summarize undesirable consequences because the articles never discuss undesirable consequences, (3) producing a nonsensical summary (a phenomenon known as text degeneration [45]), or (4) generating oversimplified summary with insufficient context (requiring decontextualization [23]). For example, the summary "They [VA] will probably make us all look like idiots." is inaccurate because it is oversimplified (does not contain enough context to stand-alone). We found that 84% (42/50) were rated as accurate, suggesting that summarizations are largely reliable. For example, one wrong summarization of an article on the Interpreter Mode of Google Translator [24] is "we will be often talking to our devices than each other. This is a bad thing." The article mentions potential mistranslation for people with thick accents but does not explicitly mention overuse. A future improvement could be allowing user feedback to enhance the potentially incorrect summaries.

[2]We avoided product-specific search terms to avoid capturing generic articles that included text, such as "Share this link on Facebook and Twitter."

Aspect Categorizing. We evaluated Blip's 10-way classification using the pilot 150 articles assigned to 10 different aspects of life. Blip's classification achieved an accuracy of 38% (F1=36%), which is not ideal but expected as the performance is comparable to other multi-class classification tasks [85] (even including those with fewer categories). Achieving higher accuracy with the zero-shot approach is difficult. In our case, the fairly complex categories (e.g., Health & Well-Being) and their potential for overlap lower the performance; misclassifications in our dataset are rarely egregious but instead happen when multiple categories could be a potential fit.

3.5 Growing Blip's Content

While our technical evaluation demonstrates the feasibility of adapting NLP techniques to extract undesirable consequences for three technology domains, Blip includes two ways for adding more undesirable consequences and additional technology domains.

First, users can click on the "Import an article" button in Blip's user interface (Figure 1-8) to add a single article via URL or PDF. Blip then runs the article through its extraction pipeline and shows a card with the summary, link, and aspect category as output. Users can assign the card to an existing technology domain or propose a new one (e.g., robotics). Added articles, cards, and technology domains are stored in a temporary database and only added after approval to avoid potential misuse and retain control over the number of technology domains that are being added.

A second option is to use Blip's bulk-import functionality, which is currently only available to developers to control the costs of accessing the OpenAI API. This functionality allows inputting a keyword (e.g., social media) or adding several URLs or an entire spreadsheet data file with several articles at once, which Blip then runs through the extraction process that can take several hours. As we described in Section 3.4, we have previously obtained online articles from a set of trusted online technology magazines by searching for, and downloading, those that discuss specific technology domains. We plan to continue using this approach for adding new domains (as we did to add more fields in our Study 2).

In addition to these two approaches for manually adding undesirable consequences, the system automatically checks for new articles in the three domains (and on the four sites listed in Table 1) on a weekly basis and adds the discussed undesirable consequences. The update frequency is flexible; we decided on weekly updates because there are usually only 3-5 new articles every week.

4 STUDY 1

Our first study evaluated Blip's overall usefulness for researchers who are experts in a technology domain that is currently covered in Blip. Specifically, our study investigated whether a catalog of undesirable consequences in Blip could enhance their awareness of potential impacts within their general CS subarea. The study design was guided by three research questions:

RQ1 Does Blip support CS researchers in discovering undesirable consequences beyond their prior knowledge and beyond searching on the internet?
RQ2 How do researchers perceive Blip's usefulness for discovering undesirable consequences in their area?
RQ3 How and when do researchers imagine using Blip during the research and development process?

4.1 Methods

We chose a within-subjects design to answer whether Blip helps researchers gain insights into undesirable consequences in their technology domain, beyond what they already know and beyond what they may discover through an online search.

4.1.1 Participants. We recruited nine computer science (CS) researchers (7 male, 2 female) aged 24-41 ($\mu = 27.55$, $\sigma = 5.66$) years old through personal connections and institutional Slack channels. Our participants are based in the US and fluent in English. To ensure participants actively worked on cutting-edge technologies, our selection criteria required participants to be CS researchers in academia or industry and develop technologies in the areas of social media (SM), voice assistants (VA), or virtual reality (VR) (3 from each) and to have authored at least one peer-reviewed publication in one of these technology domains. Importantly, our participants were not new to the topic of undesirable consequences. Two participants work on ethics-related topics (AI fairness and identifying security and privacy issues in VR). Of the 9 participants, 7 are Ph.D. students, 1 is a postdoctoral researcher, and 1 is a research scientist at a non-profit research institute. Seven participants previously worked for large multinational technology corporations relevant to their research fields. We met with 5 participants over Zoom (P3, P5-6, P8-9) and 4 in person (P1, P2, P4, P7).

4.1.2 Procedure. Each study session started by obtaining consent, followed by a brief introduction of *undesirable consequences of technology*. All participants used their own devices for the study. The recruitment email did not detail the study procedure to prevent them from preparing in advance and confounding our baseline conditions. Each participant took part in three ordered conditions:

(1) Know (**Prior Knowledge Baseline**): Participants were asked to think about and list any undesirable consequence of digital technologies in their domain of expertise (one of SM, VA, or VR), solely based on their prior knowledge.
(2) Search (**Online Search Baseline**): Participants were asked to search any online resources of their preference (e.g., Google Search, Google Scholar, Semantic Scholar) to add to the previously generated list of undesirable consequences.
(3) Blip (**Blip System**): Participants were asked to interact with Blip and list any additional consequences that they did not mention in either of the prior conditions. Participants were provided the URL for Blip without any additional instructions on how to use it.

We chose the two baseline conditions because these are plausible alternatives for exploring undesirable consequences, which technologists desire to do but not yet commonly practice [31]. Each condition was limited to 15 minutes to allow for comparable outcomes across conditions and to keep the study duration to at most one hour. None of the participants reached this limit in any of the conditions. Participants were alerted about the remaining time after 10 minutes. They were also informed that they could jump to the next condition when they could not think of, or find, more undesirable consequences. While participants were encouraged to think of as many unique undesirable consequences as possible, all of them eventually ran out of ideas and switched to the next condition.

Each study session ended with a semi-structured interview eliciting feedback on Blip. The interview asked questions about the perceived usefulness and challenges of Blip as well as how participants would integrate Blip in their research process and what future use cases they could envision for Blip.

4.1.3 Analysis. To answer whether Blip supports developers in naming undesirable consequences beyond their prior knowledge and an online search (**RQ1**), we analyzed the number of unique undesirable consequences listed during each condition. Consequences were considered as unique if they considerably differed in detail—for example, in terms of different "aspects" like privacy issues due to recording vs. data leaks or specific examples of undesirable consequences affecting certain populations in different ways. We also required consequences to be either already existing or reasonable (e.g., they could be anticipated but had to be realistic). We counted the results by the conceptual distinction made by the participant barring repetition or similar incidents (e.g., virtual reality might cause nausea or motion sickness are counted as one unique idea). Two authors independently coded and counted participants' unique consequences. During analysis, the three conditions were randomized to prevent confirmation bias towards our system, Blip. To assess the category consistency, we calculated the inter-rater reliability of the categories using Cohen's Kappa $\kappa = 85.33\%$. For the 21 responses out of 155 that the authors disagreed on, two authors discussed and decided on the final aspect. The anonymized user data can be found in the Supplementary Materials.

To answer **RQ2** and **RQ3**, we conducted a thematic analysis of the semi-structured interviews. First, two authors individually reviewed, and conducted open coding for, three interviews. Next, three authors met to consolidate and create the first draft of the codebook. Finally, two authors independently coded the remaining interviews and refined the codebook, which was discussed with the full research team. In line with prior suggestions [63], we did not calculate an inter-rater reliability score for the interview codes because our goal was to uncover more themes.

Researchers' Positionality. Our motivations and perspectives for designing, developing, and evaluating Blip are shaped by our academic and professional roles as US-based CS researchers at an R1 university. With backgrounds in HCI, NLP, Software Engineering, and Computing Ethics, we had many discussions about supporting and incentivizing researchers in various subfields of computer science to learn about and anticipate undesirable consequences of technology that all authors have experienced in the past. All of the authors have had prior research experience (as faculty, interns, or research scientists) at other universities and in industry labs, which has influenced our thinking of how the inertia to consider undesirable consequences could be overcome. While we are ultimately technologists, we see Blip not as a complete solution to the problem, but as being in a supportive role that needs to be combined with incentives and structural changes in academia.

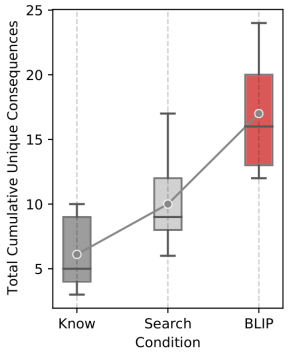

Figure 3: Cumulative unique number of consequences in each condition. The line represents means of unique consequences between the three conditions.

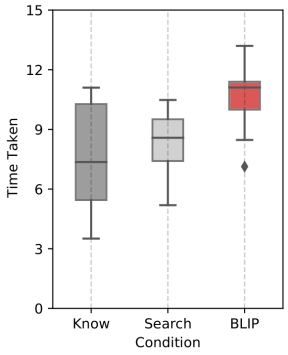

Figure 4: Total time taken for each of the three conditions separately (in minutes).

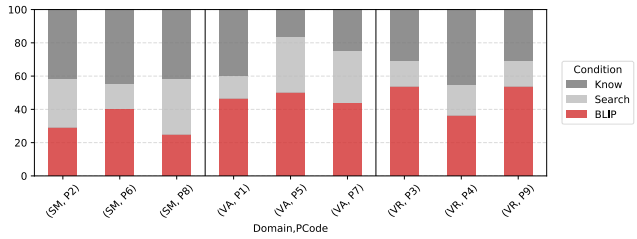

Figure 5: Percentages of total consequences by each participant for each condition of our study by technology domain.

4.2 Results

We report our findings in the following two sections.

4.2.1 ***Blip allows participants to find additional and diverse undesirable consequences (RQ1)***. Our quantitative analysis showed that Blip supported participants in discovering undesirable consequences beyond their prior knowledge and beyond what they found during an Internet search. In the Know condition, participants listed an average of 6.11 (SD=2.80) unique undesirable consequences (Figure 3). For example, participants mentioned potential implications ranging from an "echo chamber" in social media to sensitive bio-metric information used for advertisement in virtual reality. They mentioned reading about these examples in the news or research papers, in addition to sometimes citing their own work. Participants spent an average of 7.52 minutes (SD=2.68) on this first condition as shown in Figure 4. All participants switched to the next condition before the 15-minute limit was reached.

The Search condition only resulted in an average of 3.88 additional undesirable consequences (SD=1.83). Overall, participants spent more time in this condition than in the Know condition (Know: $M = 7.52$ minutes, $SD = 2.68$; Search: $M = 8.23$ minutes, $SD = 1.96$) The difference is mainly due to the time required for searching for, and reading through, information online. The fact that they found fewer ideas on average in Search makes intuitive sense because participants searched for consequences similar to or building upon what they already thought of in the Know condition. Our observations suggest that searching for resources online, such as through Google Search, Google Scholar, or Semantic Scholar, is not well-suited for finding undesirable consequence, owing to the fact that it often necessitates prior knowledge on what to search for. While participants used different combinations of search engines (5 participants only used Google Search, whereas 4 others used both Google Search and Google Scholar or Semantic Scholar), they were unable to find many new consequences. As P8 noted, *"all of the titles on Google said the same things ..., so I had to open them to see [the content], which is tiring."* P3 even had trouble finding the right keyword to search for the content, stating that *"I don't know if there just isn't so much work about [safety issues of VR] or I just searched with the wrong keywords"*. All participants switched to Blip before the 15 minutes ended, suggesting that they had exhausted ways for searching for information about undesirable consequences.

With Blip, participants were able to add an average of 7.00 undesirable consequences ($SD = 1.65$) not listed before. P3 summarized their experience by saying *"a lot of ideas just came to me that I otherwise would never have thought of."* As an expert on social media, P2 mentioned an additional 7 ideas when using Blip compared with 17 ideas in the prior baseline conditions (Know: 10, Search: 7). For instance, they commented that as social media platforms grew, governments can easily censor the population by deleting controversial topics, a consequence which they had "*totally missed*" in the previous conditions.

While participants were able to expand their list of undesirable consequences using Blip, we found that participants spent, on average, more time with Blip ($M = 10.52$ minutes, $SD = 1.85$) than in the other two conditions (see Figure 4). Our observations and analysis of post-study interviews suggest that Blip was perceived as engaging and as a tool that continuously led to new insights. For example, participants commented that Blip enables them to find examples of undesirable consequences quickly (and especially more quickly than in the Search condition). P1 mentioned that they were completely oblivious to the time they spent exploring Blip until we reminded them after 10 minutes in the Blip condition. P9 described the interface as *"addictive"* and that they would like to keep re-visiting Blip in the future. This suggests that, unlike in the Search condition, participants felt like they wanted to spend more time exploring undesirable consequences.

A detailed breakdown of the percentage of unique consequences listed in each condition by each participant can be found in Figure 5. Based on the percentage of new consequences added in each condition, Blip appears to be most useful for VR, VA, and SM in that order. Such an order could be the result of the different eras in which these technologies were introduced—the undesirable consequences of VR and VA are just becoming apparent, whereas those of SM have been known for some time and are regularly discussed in the news. Participants working on social media were therefore able to cover many consequences in the Know condition.

While the number of undesirable consequences participants were able to think of using Blip is encouraging, we also analyzed whether Blip helped participants think of more diverse instances than in the Know and Search conditions. For this analysis, we manually categorized each of their responses into one of the 10 aspects of life in Section 3.3.3. The consequences can be categorized into an

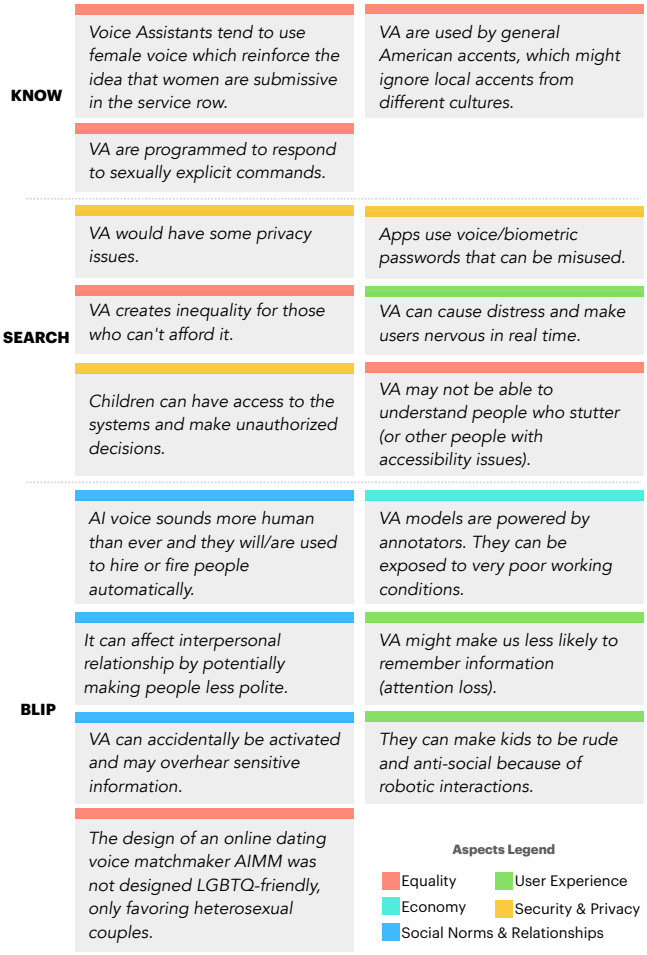

Figure 6: Quotes from P5's reporting of undesirable consequences of voice assistants and the associated aspects of life in the three conditions of our user study.

average of 4.11 aspects in the Know condition, 2.78 in the Search condition, and 5.00 in the Blip condition. To exemplify how participants broadened their list of undesirable consequences across the three conditions, Figure 6 shows that P5 first fixated on undesirable consequences related to Equality in the Know condition. Searching for consequences broadened their list to three different aspects (Security & Privacy, Equality, and User Experience). It was only when they started using Blip that they additionally thought of impacts on Economy and Social Norms & Relationships, in addition to others. As we will show in our qualitative results in the next section, several participants echoed our finding that Blip helps users diversify their list of undesirable consequences.

Altogether, our findings indicate that Blip supports users in discovering undesirable consequences beyond their prior knowledge and searching online, affirming our first research question.

4.2.2 ***How participants perceived Blip's usefulness and how it would be used (RQ2 and RQ3).*** To answer our second and third research questions, we present three high-level themes as revealed by our qualitative analysis.

Blip is useful for learning about undesirable consequences and reflecting on their own experiences. Our analysis revealed that participants found Blip improves their ability to *"think outside the box"* and made them aware of undesirable consequences that they *"had never considered"* [P3]. Many participants suggested that a tool like Blip is essential and wished it had existed earlier. For instance, P7, a Ph.D. student who studied conversational AI, said they wished they had looked through the examples or *"at least thought about these societal issues"* in their first project. They regretted working on an issue with a *"very poor understanding of the social implications."* They explained that ignoring societal ramifications can be a bigger problem for technical fields disconnected from end-users such as NLP because *"people have established benchmarks and evaluation metrics, [so] you [researchers] can work on the task without having any idea of how it's used in real life."* P1 echoed that thinking of undesirable consequence is *"difficult"* if researchers do not work specifically on fairness or accountability issues. Because Blip pre-processes the real-world examples, P7 found it to be *"easier and faster"* to learn about undesirable consequences compared to relying on their prior knowledge or online search.

Participants also stated that Blip's aspect categories are useful for broadening their knowledge of undesirable consequences. For example, P6, who authored over 5 papers on mental health issues on social media, extensively discussed undesirable consequences such as how social media can cause people to feel lonely, show *"scary"* images that can affect users' mental stage, and idealize beautified photos that make teenage girls feel bad about themselves. When searching online, they continued focusing on health-related risks such as how social media can increase depression and schizophrenia. Blip expanded their discussion to other issues on privacy, economy, online bots, limitation of freedom of speech, and interpersonal relationships. In the end, P6 commented on the benefit of having different aspect categories: *"I do feel like it's very nice to get exposed to a wide variety of topics... I feel like this would be great to anticipate [undesirable consequences of] a new social media product."*

We observed a similar pattern for P1, P3, P5, P8, and P9, all of whom focused on aspects that are related to their research areas in the baseline conditions and only included more diverse issues once they started using Blip. Note that we explicitly asked participants to list diverse undesirable consequences until they exhausted their ideas, at which point they could switch to the next condition. The fact that only three participants generated consequences across several aspects of life from the beginning suggests that many researchers and developers may fixate on specific issues. The diversity of undesirable consequences in Blip was useful for gaining insights into additional aspects and overcoming this fixation.

Participants also noted that the examples in Blip inspired them to think of their own prior experiences or additional undesirable consequences. For instance, for P1, a summary on one card (*"People have become more reliant on machines to do tasks that they are capable of doing themselves."*) inspired them to recall their own experience with voice assistants: *"[I] deliberately speak in a way that it will be easier for [the] machine to understand."* They recalled that the interaction *"fundamentally changed my behaviors."* Every participant mentioned at least one incident (*"This made me think of ..."*). As P8 said after using Blip, "*I felt that I was exposed to a bunch of possible directions that you can see. I feel that I'm learning a lot.*"

Blip can be useful for brainstorming undesirable consequences, writing ethics statements, and for different stakeholders. We found that participants appreciated Blip for helping them brainstorm undesirable consequences of their own innovations, including when writing an ethics statement or introduction for a paper. P4, who studied security issues on VR devices, suggested that Blip could be useful for finding arguments and citations for their publication discussing the undesirable consequence of VR. P1-2, P4, and P7-8 all mentioned that they perceive Blip to be useful to get ideas for ethics statements or the introduction of their papers. P4 was hopeful that Blip could play a role in establishing a *"brainstorming phrase"* for CS researchers to determine the research questions to address. Similarly, P7 indicated that they would use a system like Blip to think about undesirable consequences of a new topic early on, stating that *"it will be very useful for brainstorming and will [help me] process a lot of information faster."*

Participants also felt that Blip could be useful for a variety of stakeholders, not just for themselves. Several participants suggested that it would provide a good introduction to undesirable consequences for the general public, researchers, or developers who are new to a particular field. As P9 mentioned: *"I think in order to get a sense of the full paradigm of VR, you have to have a tool like this because you can't just read a bunch of disconnected articles about this."* P1 made a similar comment that users could quickly browse through a *"broad spectrum of issues"*. P3 also appreciated that Blip can enable readers to *"quickly skim through the summaries"* without delving into each online resource.

In summary, the results of our first study suggest that Blip supports the discovery of more and more diverse undesirable consequences relevant to specific technology domains compared to prior knowledge and an online search (**RQ1**). Blip was perceived useful to learn about about undesirable consequences and reflect on their own experience (**RQ2**). In addition, participants found that Blip can be useful for brainstorming undesirable consequences, writing ethics statements, and for different stakeholders (**RQ3**).

4.2.3 Participant Feedback and System Changes. Our interviews revealed several opportunities for improving Blip, which we subsequently implemented. Specifically, participants (P2, P4-5) suggested that in addition to online articles, it would be helpful to also have academic articles available—in particular those that uncover and discuss undesirable consequences. We therefore added the functionality to parse and summarize academic papers, adding the last twenty years of papers from the CHI proceedings (2003-2023) as a data source. In Blip's GUI, users can filter whether they want to see all data sources, only academic papers, or only articles from online magazines. We also made minor changes to the GUI based on their feedback, such as changing the appearance of the summaries and buttons. Additionally, we included the Llama2 open-source model in the backend as more large language models become available [105].

5 STUDY 2

Our first study showed that Blip can indeed increase researchers' awareness of undesirable consequences in their CS subfield compared to relying on their prior knowledge or searching online. What remained unanswered was whether Blip is useful for collecting, considering, and acting on undesirable consequences relevant to specific projects users work on. Our second study, therefore, focuses on the following research question (**RQ4**): To what extent is Blip useful and actionable for users' *own projects*? We study this question both objectively (i.e., whether they can find undesirable consequences that are relevant to their projects) and subjectively (whether they perceive Blip as useful and actionable).

5.1 Methods

5.1.1 Participants. We recruited six CS researchers (4 male, 2 female) aged 23-26 ($\mu = 25.00$, $\sigma = 2.00$) years old through personal connections and institutional Slack channels. Our participants are based in the US and fluent in English and all are currently Ph.D. students with experience in the technology industry through internships or prior work experience. The six researchers work in Computer Vision, Vision Language Models, Mobile Technology, Computational Biology, Robotics, and Ubiquitous Computing. To protect their anonymity, we only refer to their general research directions and avoid discussing the specifics of any project.

5.1.2 Procedure. We met participants over Zoom (P1-2, 4-5) and in person (P3, 6). Before each session, we used Blip's bulk-import functionality (see Section 3.5) to filter, summarize, and categorize new content specific to the six CS subfields that participants' work contributes to. We added all content on Blip, including a list of domain keywords and the imported articles, before the user studies to ensure that participants could explore consequences in their own subfields (e.g., Ubiquitous Computing). At the beginning of the study, we asked participants to describe their current project. They were then explicitly instructed to use Blip to bookmark as many undesirable consequence as they may find relevant to their own projects as if they were trying to establish a comprehensive list. Participants could open the articles in Blip if they were interested. To approximate a real-world usage scenario and avoid making participants feel observed, the experimenter left the study session until participants messaged the authors that they had gained a sufficient overview of the relevant undesirable consequence. Participants were told that they had a maximum of 30 minutes to explore Blip.

At the end of the session, we conducted a brief interview with each participant. Specifically, we asked participants to share their screens and explain they found the bookmarked articles relevant to their projects. This was done to understand how participants would use the undesirable consequences described in the bookmarked articles and whether they were actionable. We also asked participants which features in Blip they liked and what may encourage or prevent them from using Blip, as well as what improvements they would recommend. After completing the session, participants were sent a link to a post-study survey asking them about the usefulness of Blip for considering undesirable consequences in their own projects. All responses were anonymous to reduce potential response bias. (See Figure 7 for specific questions.)

5.1.3 Measures. We collected the total time participants used Blip, the number and content of bookmarked articles, and responses to the survey questions. We recorded the post-study interviews for qualitative analysis. Similar to Study 1, two authors individually reviewed and conducted open coding on the initial two interviews. The two authors then independently coded several more interviews

and refined the codebook. We did not calculate the inter-rater reliability for the codes to discover more emergent themes [63].

5.2 Results

Participants spent an average of 13.11 minutes (SD = 4.72) browsing through Blip and bookmarked an average of 7.67 articles (SD = 3.94, Min = 5, Max = 13) relevant to their research articles (See Supplementary Materials for examples of bookmarked articles). The result suggests that participants can find undesirable consequences through Blip that are relevant to their own projects and that they can do so in less than 15 minutes.

Our post-study survey results showed that participants perceived Blip generally as useful and actionable (see Fig. 7). Specifically, four of our six participants agreed or strongly agreed that Blip is useful for learning and considering undesirable consequences relevant to their own project (two were neutral). All participants agreed or strongly agreed that Blip would be useful for others working in their area. Blip was also seen as an inspiration to think of undesirable consequences of participants' projects (five agreed, one strongly agreed). Finally, three participants agreed or strongly agreed that Blip provides them with new perspectives for their project, suggesting its actionability (two were neutral) and three participants would use it for future projects (three were neutral).

5.2.1 Interview Results. The interviews with participants helped to shed further light on these results. First, **participants found and bookmarked summaries that were relevant to their projects**. For example, P6 who worked on ubiquitous computing, bookmarked a mixture of online articles and CHI papers. Their current project was focused on building smart electrocardiogram (ECG) systems, so they bookmarked summaries that were particularly relevant to ECG and wearables. One summary suggested that wrist devices lack accuracy when people are mobile or if the tasks require physical activity; another one described ECG biometrics as lacking robustness when deployed in the wild. During the interview, when asked about the relevance of the bookmarked summaries for their project, they pointed out an article titled *What to Put on the User: Sensing Technologies for Studies and Physiology Aware Systems*. P6 said that *"the source seems to cover challenges and the different physiological signals that we can use for wearable technologies [...] I think it will be helpful to know [the signals] besides what I'm using right now and be aware of the issues in the future."* P6 suggested Blip made them look at potential real-world implications, instead of merely finding, replicating, and improving *"lower-level and technical systems from the very beginning."* They commented that Blip offers a very different perspective to think of their research in ubiquitous computing. They later notified the authors that they reviewed the bookmarked list after our study and read the articles in detail.

Similarly, P2, who works on vision language models, bookmarked an article describing that voice assistants may have the potential to reduce creativity due to overpersonalization. When asked about this article in the interview, P2 mentioned that their research project includes a *"component to add user's preference and context input to improve the model"* but they did not think that this would have any consequence at all. P2 continued that *"overpersonalization can be a problem, right? Maybe we [our team] should think more about this feature."* P4, who worked on a computational biology project and *"did not think about ethics too much,"* commented that they often focus on the technical aspect but had never read so many perspectives about potential *"risks"* in the real world before using Blip. They also pointed to articles that they thought relevant to their projects about the undesirable consequence of gene therapy after searching keywords such as "gene therapy" and "t-cell receptors".

Second, the interviews showed that **participants generally perceived Blip as useful and actionable**. For example, P5 mentioned that they often read magazines such as The Verge, and *"[using] Blip feels like reading a more relevant news page."* P5 mentioned that Blip helps them to rethink ways to alleviate the potential consequences of robots. They found an article about the impact of robots on economic inequality and commented that they had not thought much about this issue since automation seems always *"the way to go."* They had also bookmarked two articles about potential solutions titled *"One way to get self-driving cars on the road faster: let insurers control them"* and *"EU proposals would classify robots as electronic persons."* Inspired by the two articles, P5 proposed how one could address the problem, suggesting they could implement *"subscription service where robots just act as an agent so that everyone can control the robots and receive the payment."* P5 acknowledged that this *"of course, is a very naive solution,"* but *"I should start think more about these issues"* as Blip have exposed many articles to them.

Third, the interviews explained why **some, but not all, participants thought they would integrate Blip into their work in the future**. Several participants stressed the importance of having Blip. For example, P2 stated *"I think using this tool is particularly useful for ethical consideration, like [the] ethical statements section in the ACL papers because most of [the] time researchers don't know what concrete examples to write for those sections, and this provides good resources."* P2 continued to say that *"a lot of the resources are from trendy topics on magazines and the venues that I'm not familiar with, [un]like ACL or other pure machine learning conferences."* P2 later told the authors that they would try to use Blip for their next paper submission. P1 also mentioned that they would use Blip to find relevant literature, which they would like to include in the introduction or related work sections. P3 also stressed having such a catalog of consequences is *"necessary"* to keep up with the related work and how the media portrays some issues. According to P3, this is in contrast to Google Scholar feed which only recommends technical work for them.

Other participants who did not think they would use Blip in the future (P4-6) viewed the task of addressing undesirable consequences as tangential to their primary focus on technical development. Their comments suggest that a tool alone will not change people's perceived responsibility for actively considering undesirable consequences. For example, P6 mentioned that they learned *"a lot [about] social impact"* by looking at Blip, but their research *"primarily focuses on building a technical system."* They were not sure whether they should be the person thinking about undesirable consequences. P5 mentioned that most of their research considered *"achieving human-level dexterity of robotics,"* but most of the undesirable consequences in Blip are very *"socio-economical."* They did not think that they are at the right position or qualified to think about and address these issues.

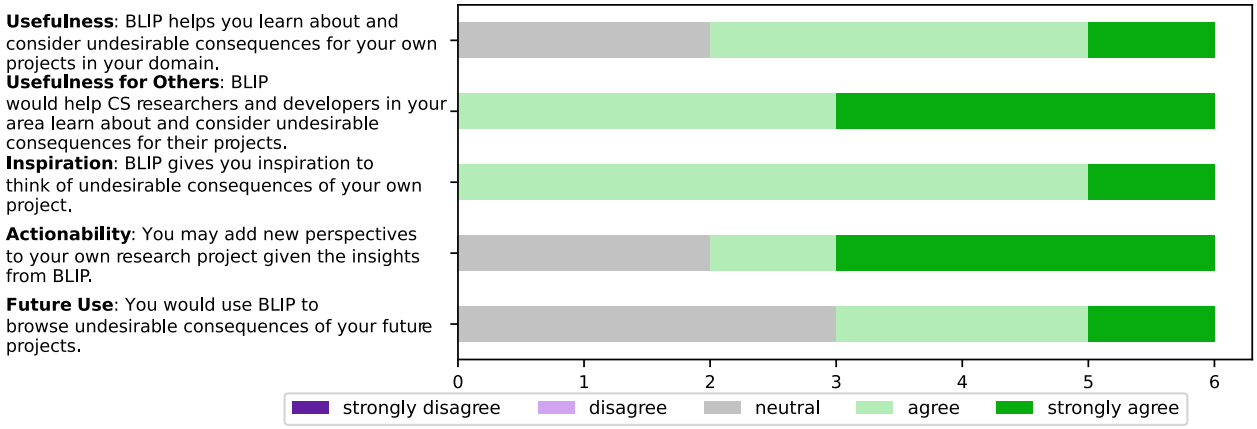

Figure 7: Post-study Survey Responses in Study 2. The x-axis shows the number of participants (N=6)

Finally, **participants suggested opportunities for making Blip more useful for them**. As in the previous study, several participants suggested that expanding Blip's catalog to other sources may provide additional inspiration. For example, P3 mentioned they found *"some cool articles"* and relevant CHI papers, but they were interested in *"how many papers in my field [Mobile Technology] and outside CHI revenue portrayed these problems."* They suggested including a list of publication venues and workshops across disciplines but realized that it probably went beyond the Blip's scope.

P5 commented on improving Blip's actionability by suggesting that the system would greatly benefit researchers if it could include a *"solution"* to the undesirable consequences. In a similar vein, P1 suggested adding potential expert reviews or opinions. *"I'm working on a very specific and new domain. I've talked about undesirable consequence with my labmates but I'd be interested in hearing more experts to comment on their thoughts."*

In summary, the interviews revealed that participants recognized the Blip's potential and gained ideas of potential consequences that they would need to address in their projects. Participants were also envisioning Blip 2.0, suggesting the need for tools that summarize the broad range of scientific literature from diverse fields on undesirable consequences and aid in finding solutions. However, our results also showed that the onus of proactively addressing undesirable consequence cannot rest solely on Blip; integrating into existing research processes could further amplify its impact.

6 DISCUSSION

Our goal in this work was to evaluate whether providing CS researchers with an easily accessible catalog of undesirable consequences of digital technologies could improve their ability to learn about and consider adverse effects, as prior work had suggested [31, 64, 98]. To study this question, we developed Blip, a web-based prototype that leverages language models to automatically derive undesirable consequences from any given online article. Blip addresses the difficulty of having a broad knowledge of potential undesirable consequences, which, according to Merton, is "the most obvious limitation to a correct anticipation." [65]

Our results show Blip's potential for supporting CS researchers in gaining awareness of a broad range of undesirable consequences. In Study 1, Blip supported researchers in finding more, and more diverse, undesirable consequences of technology in their CS subdiscipline even after listing ones from prior knowledge and after searching online. When relying on their prior knowledge, we found that participants only thought of an average of 6.11 unique undesirable consequences despite being experts in their technology domain. They were often stuck describing undesirable consequences within one or two commonly known "aspects" (that were sometimes part of their research focus). This indicates that many researchers do not have a thorough and broad awareness of the undesirable consequences within their technology domain. Intuitively, searching online might be a better option to explore additional undesirable consequences to extend users' prior knowledge. However, our results showed that searching online only added an average of 3.88 undesirable consequences and was perceived as a tedious approach. The fixation issue persisted when searching online, that is, participants mostly used search terms related to the undesirable consequences they had already listed. This limitation underscores the insufficiency of traditional methods, as they often lead to a narrow focus, overlooking broader and potentially more impactful consequences. Compared to the two baseline conditions, Blip was able to support participants in listing and learning about undesirable consequences that were often beyond the commonly known ones. To summarize these results, our study demonstrates that relying on prior knowledge and an online search, without any tooling support, is often perceived by the participants as tedious and insufficient for thinking about the undesirable consequences of technology.

In Study 1, the qualitative responses further illustrate the most helpful parts of Blip's design. We found that participants perceived Blip's summaries of undesirable consequences as beneficial for efficiently gaining an overview, while appreciating having access to the original articles to ensure information integrity. Blip's categorization of different life aspects was seen as a motivating nudge for exploring consequences broadly—a finding that is in line with the results of our quantitative analysis. The result extends prior work in creativity and cognition that has found that providing a solution space using a set of dimensions breaks people's tendency to fixate [74, 96]: Providing a diverse set of undesirable consequences

can help technology experts consider societal implications broadly and reveal those that they would have otherwise not thought about.

Study 2 aimed to evaluate whether Blip provides actionable information when freely using it to find potential undesirable consequences relevant to researchers' specific projects, rather than to the whole field. We found that participants took less than 15 minutes, on average, to gather a set of consequences relevant to a specific research project and bookmarked an average of 7.67 unique undesirable consequences during this time. While this second study was not designed to determine whether participants were able to *comprehensively* find undesirable consequences, participants' comments suggested that the ones they found inspired them to think of undesirable consequences more broadly. The finding also suggests that Blip could be a useful resource for gathering undesirable consequences to include in a paper's ethics statement, well beyond the average of 0.6 words that are included in ethics statements in NeurIPS AI papers, for example [6, 71].

Our follow-up interview and survey, however, painted a more complicated picture of anticipating undesirable consequences using Blip in practice. After using Blip for their own projects, two of six participants in the second study were neutral on Blip's usefulness for their projects, though all agreed that the system is useful *for others*. Some participants acknowledged that they are not in the right position to think about these issues, though they learned "a lot about social impact." The result resonates with the narrative reflected in Do et al.'s work [31] that CS researchers tend to deflect the responsibility to consider the adverse effects of technology.

This brings us to how we see Blip can support researchers in the future. Participants suggested that they would use Blip for inspiration when writing broader impact statements in papers, which could lower the perceived burden of writing them [2] and potentially counteract the focus on desirable outcomes [103]. However, ultimately it would be ideal if CS researchers routinely learn about, anticipate, and reflect on undesirable consequences—as has been repeatedly advocated for [44, 61]—and that they do so early and proactively when addressing undesirable consequences is still feasible [31]. We envision Blip as a tool that supports researchers in doing so, both by appealing to their intrinsic curiosity and by having extrinsic incentives such as fulfilling the requirements of a conference, grant agency, and institution.

Per Study 2, Blip, while a useful tool, is not a magical bullet to achieve this paradigm alone. Doing so would not only require a catalog of concrete consequences by Blip but also a systematic change in culture and structural incentives. Tools like Blip may spark new conversations and alleviate the perceived burden of anticipating undesirable consequences particularly if research institutions evolve to actively encourage this reflection. For example, researchers can easily explore a wide variety of undesirable consequences in their domain for past incidents before launching a new project. When writing ethics statements (e.g., for papers or grant proposals), researchers may use Blip to efficiently and thoroughly examine their case. They could engage with the information and stay updated on the latest undesirable consequences. A crucial aspect of this institutional change is nurturing a mindset among researchers that recognizes the importance of contemplating these challenges, thereby embedding the practice of considering undesirable consequences as a fundamental aspect of responsible research.

In the long run, we envision Blip to become an integral part of the technology development and research process by using strong incentives and implementing systemic changes as suggested in prior work [10, 44]. We hope that using Blip will inspire the research community to work towards such a future in which learning about and anticipating undesirable consequences soon becomes the norm.

7 LIMITATION & FUTURE WORK

A limitation of the Blip prototype is that it currently only uses online technology magazines and CHI papers to retrieve undesirable consequences of technology. These may not be a comprehensive catalog of undesirable consequences. In particular, the catalog may not adequately reflect the consequences that diverse user groups experience given that these articles are commonly written for 'tech-savvy' audiences. In future work, we plan to systematically explore the difference in the reporting of undesirable consequences across tech magazines, newspapers, and research papers from diverse fields and augment Blip's sources. Future work should also incorporate non-English language articles, or non-American media outlets, to better reflect the effect of technology on diverse users. Another improvement is to incorporate multiple aspects rather than one category. Encouraged by the feedback from our participants, we also believe that there are exciting opportunities to enable citizen scientists to document their personal experiences with undesirable consequences in Blip. This could satisfy users' desire to share their own experiences while enabling insights into potential differential effects of technology on people.

We designed Blip using LLMs due to the increasing performance on NLP tasks such as classification and summarization [15]. Nevertheless, LLMs can also introduce serious undesirable consequences, such as model hallucination, biases in the training data, and a limited understanding of emerging fields. 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. In this paper, we offered a preliminary evaluation of the individual components within Blip, using quantitative metrics such as F-1 score. While our metrics are comparable to similar ML tasks (see Section 3.4), there is a tradeoff between foraging consequences at scale and ensuring perfect accuracy. Involving citizen scientists could help improve the overall quality of the Blip data curation process, such as by providing their feedback on the article relevance and label accuracy and by creating their summary and labels on the existing articles.

We also foresee several different use cases for Blip. For example, our participants indicated that Blip could help researchers seek inspiration when writing broader impact statements for publications or grant proposals. Researchers may use Blip to introduce the background knowledge in their areas to highlight their solutions to the public. Additionally, policymakers and practitioners could use it to inform their work. Journalists could leverage a system like Blip to discover new angles when reporting news, especially technology mishaps (see a related tool specifically designed for journalists for story inspiration [59]). The interested public can fairly efficiently learn about how digital technology has already affected, and will continue to affect, their lives. We believe that Blip could also aid different stakeholders in reflecting on societal issues.

ACKNOWLEDGMENTS

We thank our participants and the anonymous reviewers for their valuable feedback. We also thank Sandy Kaplan, Alicia Gao, Kevin Feng, and Yadi Wang for their helpful suggestions. This work was funded by the National Science Foundation as part of the awards IIS-2006104, ER2-2315937, and SMA-2315937.

REFERENCES

[1] IUI Program Chairs 2024. [n. d.]. Reflecting on Societal Implications of IUI Research. https://iui.acm.org/2024/societal_impact.html, last accessed: February 27, 2024.

[2] Grace Abuhamad and Claudel Rheault. 2020. Like a Researcher Stating Broader Impact For the Very First Time. arXiv:2011.13032 [cs.CY]

[3] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[5] Joan S Ash, Dean F Sittig, Richard H Dykstra, Kenneth Guappone, James D Carpenter, and Veena Seshadri. 2007. Categorizing the unintended sociotechnical consequences of computerized provider order entry. *International journal of medical informatics* 76 (2007), S21–S27.

[6] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. AI Ethics Statements: Analysis and Lessons Learnt from NeurIPS Broader Impact Statements. 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, 2047–2056. https://doi.org/10.1145/3531146.3533780

[7] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* 30, 5, Article 74 (sep 2023), 38 pages. https://doi.org/10.1145/3589955

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Kr. Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* 63 (2019), 4:1–4:15.

[9] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[10] Michael S. Bernstein, Margaret Levi, David Magnus, Betsy A. Rajala, Debra Satz, and Quinn Waeiss. 2021. Ethics and society review: Ethics reflection as a precondition to research funding. *Proceedings of the National Academy of Sciences* 118, 52 (2021), e2117261118. https://doi.org/10.1073/pnas.2117261118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2117261118

[11] A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. 2021. *Introducing the NeurIPS 2021 Paper Checklist.* https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b

[12] Alan Borning, Batya Friedman, and Nick Logler. 2020. The'invisible'materiality of information technology. *Commun. ACM* 63, 6 (2020), 57–64.

[13] 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

[14] 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

[15] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[16] Zana Bu cinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. arXiv:2306.03280 [cs.HC]

[17] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[18] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. 2017. Unintended Consequences of Machine Learning in Medicine. *JAMA* 318, 6 (08 2017), 517–518. https://doi.org/10.1001/jama.2017.7797

[19] A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. Chau. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE Computer Society, Los Alamitos, CA, USA, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

[20] Comms Chairs. 2021. NeurIPS 2021 Ethics Guidelines. https://blog.neurips.cc/2021/08/23/neurips-2021-ethics-guidelines/

[21] An-Chin Cheng, Chung-Jen Chen, and Chia-Yon Chen. 2008. A fuzzy multiple criteria comparison of technology forecasting methods for predicting the new materials development. *Technological Forecasting and Social Change* 75 (2008), 131–141.

[22] Shruthi Sai Chivukula, Ziqing Li, Anne C. Pivonka, Jingning Chen, and Colin M. Gray. 2022. Surveying the Landscape of Ethics-Focused Design Methods. arXiv:2102.08909 [cs.HC]

[23] Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics* 9 (2021), 447–461.

[24] Julian Chokkattu. 2019. Google assistant can now translate speech through your phone. https://www.wired.com/story/google-assistant-can-now-translate-on-your-phone/

[25] Simone Cunningham, Chloe C Hudson, and Kate Harkness. 2021. Social media and depression symptoms: a meta-analysis. *Research on child and adolescent psychopathology* 49, 2 (2021), 241–253.

[26] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5636–5646. https://doi.org/10.18653/v1/D19-1565

[27] Frank De Zwart. 2015. Unintended but not unanticipated consequences. *Theory and Society* 44, 3 (2015), 283–297.

[28] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.

[29] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-Based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 623–627. https://doi.org/10.1145/3527927.3535197

[30] Tawanna Dillahunt, Airi Lampinen, Jacki O'Neill, Loren Terveen, and Cory Kendrick. 2016. Does the Sharing Economy do any Good?. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion* (San Francisco, California, USA) *(CSCW '16 Companion)*. Association for Computing Machinery, New York, NY, USA, 197–200. https://doi.org/10.1145/2818052.2893362

[31] 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

[32] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 172–186. https://proceedings.mlr.press/v81/ekstrand18b.html

[33] J. Morley English and Gerard L. Kernan. 1976. THE PREDICTION OF AIR TRAVEL AND AIRCRAFT TECHNOLOGY TO THE YEAR 2000 USING THE DELPHI METHOD. *Transportation Research* 10 (1976), 1–8.

[34] Felix Anand Epp, Tim Moesgen, Antti Salovaara, Emmi Pouta, and undefined-dil Gaziulusoy. 2022. Reinventing the Wheel: The Future Ripples Method for

Activating Anticipatory Capacities in Innovation Teams. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Virtual Event</city>, <country>Australia</country>, </conf-loc>) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 387–399. https://doi.org/10.1145/3532106.3534570

[35] Karën Fort, Min Yen Kan, and Yulia Tsvetkov. 2021. ACL establishes its Ethics Committee. https://www.aclweb.org/portal/content/acl-establishes-its-ethics-committee

[36] Batya Friedman and David Hendry. 2012. The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1145–1148. https://doi.org/10.1145/2207676.2208562

[37] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. *Value Sensitive Design and Information Systems*. Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

[38] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[39] Jerry Glenn. 1972. Futurizing Teaching vs. Futures Courses. *Social Science Record* (1972). https://api.semanticscholar.org/CorpusID:141272607

[40] Colin M. Gray and Shruthi Sai Chivukula. 2019. Ethical Mediation in UX Practice. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300408

[41] Artefact Group. 2017. The Tarot Cards of Tech. last accessed February 27, 2024.

[42] Karen Hao. 2021. Ai voice actors sound more human than ever-and they're ready to hire. https://www.technologyreview.com/2021/07/09/1028140/ai-voice-actors-sound-human/

[43] Michael I Harrison, Ross Koppel, and Shirly Bar-Lev. 2007. Unintended consequences of information technologies in health care—an interactive sociotechnical analysis. *Journal of the American medical informatics Association* 14, 5 (2007), 542–549.

[44] Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. 2021. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. arXiv:2112.09544 [cs.CY]

[45] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=rygGQyrFvH

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

[47] Anna Iovine. 2021. Amazon Alexa told a 10-year-old to plug a charger into electrical outlet. https://mashable.com/article/alexa-wall-electrical-outlet-challenge-mistake

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

[49] René Just, Darioush Jalali, and Michael D. Ernst. 2014. Defects4J: a database of existing faults to enable controlled testing studies for Java programs. In *Proceedings of the 2014 International Symposium on Software Testing and Analysis* (San Jose, CA, USA) *(ISSTA 2014)*. Association for Computing Machinery, New York, NY, USA, 437–440. https://doi.org/10.1145/2610384.2628055

[50] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 94, 15 pages. https://doi.org/10.1145/3526113.3545660

[51] Hyeonsu B. Kang, Gabriel Amoako, Neil Sengupta, and Steven P. Dow. 2018. Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174180

[52] Mariska Kleemans, Serena Daalmans, Ilana Carbaat, and Doeschka Anschütz. 2018. Picture perfect: The direct effect of manipulated Instagram photos on body image in adolescent girls. *Media Psychology* 21, 1 (2018), 93–110.

[53] Amy J Ko, Alannah Oleson, Neil Ryan, Yim Register, Benjamin Xie, Mina Tari, Matthew Davidson, Stefania Druga, and Dastyni Loksa. 2020. It is time for more critical CS education. *Commun. ACM* 63, 11 (2020), 31–33.

[54] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. https://doi.org/10.1145/3526113.3545693

[55] Bilal Kılı c and Semih Soran. 2019. How Can an Ab-Initio Pilot Avert a Future Disaster: A Pedagogical Approach to Reduce The Likelihood of Future Failure. *Journal of Aviation* 3, 1 (2019), 1 – 14. https://doi.org/10.30518/jav.508336

[56] Joseph Lindley, Paul Coulton, and Miriam Sturdee. 2017. Implications for Adoption. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 265–277. https://doi.org/10.1145/3025453.3025742

[57] Xuguang Liu, Camille B Carroll, Shou-Yan Wang, John Zajicek, and Peter G Bain. 2005. Quantifying drug-induced dyskinesias in the arms using digitised spiral-drawing tasks. *Journal of Neuroscience Methods* 144, 1 (May 2005), 47–52.

[58] 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/

[59] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the News: Digital Creativity Support for Journalists *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174049

[60] Joseph P Martino. 2003. A review of selected recent advances in technological forecasting. *Technological forecasting and social change* 70, 8 (2003), 719–733.

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

[62] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[63] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[64] Sean McGregor. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

[65] Robert K Merton. 1936. The Unanticipated Consequences of Purposive Social Action. *American sociological review* 1, 6 (1936), 894–904.

[66] Rachel Metz. 2020. The nauseating disappointment of Oculus rift. https://www.technologyreview.com/2016/05/05/245975/the-nauseating-disappointment-of-oculus-rift/

[67] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[68] Saif Mohammad. 2022. Ethics Sheets for AI Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8368–8379. https://doi.org/10.18653/v1/2022.acl-long.573

[69] Carol Moser, Sarita Y. Schoenebeck, and Katharina Reinecke. 2016. Technology at the Table: Attitudes About Mobile Phone Use at Mealtimes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (Santa Clara, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 1881–1892. https://doi.org/10.1145/2858036.2858357

[70] Priyanka Nanayakkara, Nicholas Diakopoulos, and Jessica Hullman. 2020. Anticipatory Ethics and the Role of Uncertainty. arXiv:2011.13170 [cs.CY]

[71] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 795–806. https://doi.org/10.1145/3461702.3462608

[72] Atsuko Natatsuka, Ryo Iijima, Takuya Watanabe, Mitsuaki Akiyama, Tetsuya Sakai, and Tatsuya Mori. 2019. Poster: A First Look at the Privacy Risks of Voice Assistant Apps. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2633–2635. https://doi.org/10.1145/3319535.3363274

[73] Peter G Neumann. 2008. The Risks Digest. *The Risks Digest* (2008). http://catless.ncl.ac.uk/Risks/

[74] Tricia J. Ngoon, C. Ailie Fraser, Ariel S. Weingarten, Mira Dontcheva, and Scott Klemmer. 2018. Interactive Guidance Techniques for Improving Creative Feedback. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173629

[75] Tricia J. Ngoon, Joy O Kim, and Scott Klemmer. 2021. Shöwn: Adaptive Conceptual Guidance Aids Example Use in Creative Tasks. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS*

'21). Association for Computing Machinery, New York, NY, USA, 1834–1845. https://doi.org/10.1145/3461778.3462072

[76] John Nworie. 2011. Using the Delphi Technique in Educational Technology Research. *TechTrends* 55 (2011), 24–30.

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

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

[79] Rock Yuren Pang, Jack Cenatempo, Franklyn Graham, Bridgette Kuehn, Maddy Whisenant, Portia Botchway, Katie Stone Perez, and Allison Koenecke. 2023. Auditing Cross-Cultural Consistency of Human-Annotated Labels for Recommendation Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (, Chicago, IL, USA,) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1531–1552. https://doi.org/10.1145/3593013.3594098

[80] 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]

[81] Rock Yuren Pang and Katharina Reinecke. 2023. Anticipating Unintended Consequences of Technology Using Insights from Creativity Support Tools. arXiv:2304.05687 [cs.HC]

[82] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. https://doi.org/10.1145/3526113.3545616

[83] Nassim Parvin and Anne Pollock. 2020. Unintended by Design: On the Political Uses of "Unintended Consequences". *Engaging Science, Technology, and Society* 6 (2020), 320–327.

[84] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. 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 225, 16 pages. https://doi.org/10.1145/3544548.3580907

[85] Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2289–2299. https://doi.org/10.18653/v1/P18-1213

[86] Katharina Reinecke and Abraham Bernstein. 2011. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Trans. Comput.-Hum. Interact.* 18, 2, Article 8 (jul 2011), 29 pages. https://doi.org/10.1145/1970378.1970382

[87] Bradford W Reyns, Melissa W Burek, Billy Henson, and Bonnie S Fisher. 2013. The unintended consequences of digital technology: Exploring the relationship between sexting and cybervictimization. *Journal of Crime and Justice* 36, 1 (2013), 1–17.

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

[89] Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2182–2194. https://doi.org/10.18653/v1/2021.acl-long.170

[90] Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, D M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8 (1976), 382–439.

[91] Tate Ryan-Mosley. 2021. How digital beauty filters perpetuate colorism. https://www.technologyreview.com/2021/08/15/1031804/digital-beauty-filters-photoshop-photo-editing-colorism-racism

[92] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9080–9102. https://doi.org/10.18653/v1/2023.acl-long.505

[93] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[94] Hong Shen, Wesley H. Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 850–861. https://doi.org/10.1145/3442188.3445971

[95] Taeyoung Shin. 1998. Using Delphi for a Long-Range Technology Forecasting, and Assessing Directions of Future R&D Activities The Korean Exercise. *Technological Forecasting and Social Change* 58 (1998), 125–154.

[96] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-Powered Real-time Semantic Modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 609–624. https://doi.org/10.1145/2984511.2984578

[97] Pao Siangliulue, Joel Chan, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition* (Glasgow, United Kingdom) *(C&C '15)*. ACM, New York, NY, USA, 83–92. https://doi.org/10.1145/2757226.2757230

[98] Michael Skirpan, Nathan Beard, Srinjita Bhaduri, Casey Fiesler, and Tom Yeh. 2018. Ethics Education in Context: A Case Study of Novel Ethics Activities for the CS Classroom. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) *(SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 940–945. https://doi.org/10.1145/3159450.3159573

[99] Madhulika Srikumar, Rebecca Finlay, Grace Abuhamad, Carolyn Ashurst, Rosie Campbell, Emily Campbell-Ratcliffe, Hudson Hongo, Sara R Jordan, Joseph Lindley, Aviv Ovadya, et al. 2022. Advancing ethics review practices in AI research. *Nature Machine Intelligence* 4, 12 (2022), 1061–1064.

[100] Kate Starbird. 2019. Disinformation's spread: bots, trolls and all of us. *Nature* 571, 449 (2019).

[101] Amanda Stent. 2022. *Guidelines for ethics reviewing.* https://aclrollingreview.org/ethicsreviewertutorial

[102] Miriam Sturdee, Joseph Lindley, Conor Linehan, Chris Elsden, Neha Kumar, Tawanna Dillahunt, Regan Mandryk, and John Vines. 2021. Consequences, Schmonsequences! Considering the Future as Part of Publication and Peer Review in Computing Research. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 95, 4 pages. https://doi.org/10.1145/3411763.3441330

[103] Karl-Erik Sveiby, Beata Ulrica Segercrantz, Pernilla Gripenberg, Andreas Eriksson, and Alexander Aminoff. 2009. Unintended and Undesirable consequences of Innovation. In *Proceedings of the XX ISPIM Conference*, K.R.E. Huizingh, S. Conn, M. Torkkeli, and I. Bitran (Eds.). The XX ISPIM Conference - The Future of Innovation ; Conference date: 21-06-2009 Through 24-06-2009.

[104] Google People + AI Research Team. 2021. People + AI guidebook. https://pair.withgoogle.com/guidebook/

[105] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[106] Kentaro Toyama. 2015. *Geek heresy: Rescuing social change from the cult of technology.* PublicAffairs.

[107] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).

[108] T.B. Ward. 1994. Structured Imagination: the Role of Category Structure in Exemplar Generation. *Cognitive Psychology* 27, 1 (1994), 1–40. https://doi.org/10.1006/cogp.1994.1010

[109] W Timothy Weaver. 1971. The Delphi forecasting method. *The Phi Delta Kappan* 52, 5 (1971), 267–271.

[110] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652 [cs.CL]

[111] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda B. Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization & Computer Graphics* 26, 01 (jan 2020), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

[112] 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.

[113] Xiaotong (Tone) Xu, Rosaleen Xiong, Boyang Wang, David Min, and Steven P. Dow. 2021. IdeateRelate: An Examples Gallery That Helps Creators Explore Ideas in Relation to Their Own. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 352 (oct 2021), 18 pages. https://doi.org/10.1145/3479496

[114] Roman V Yampolskiy. 2019. Predicting future AI failures from historic examples. *foresight* 21, 1 (2019), 138–152.

[115] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2082–2096. https://doi.org/10.1145/2998181.2998235

[116] Enhao Zhang and Nikola Banovic. 2021. Method for Exploring Generative Adversarial Networks (GANs) via Automatically Generated Image Galleries *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 76, 15 pages. https://doi.org/10.1145/3411764.3445714

[117] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics* 12 (2023), 39–57. https://api.semanticscholar.org/CorpusID:256416014

[118] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (nov 2018), 23 pages. https://doi.org/10.1145/3274463