Title

8% - 10% of algorithmic recommendations are ‘bad’, but… an exploratory risk-utility meta-analysis and its regulatory implications

Authors

Hilbert, Martin
Thakur, Arti
Ji, Feng
et al.

Publication Date

2023-09-11

# 8% - 10% of algorithmic recommendations are 'bad', but…
## an exploratory risk-utility meta-analysis and its regulatory implications

Authors' version 11/29/2023

This article is being published as:

Hilbert, M., Thakur, A., Flores, P. M., Zhang, X., Bhan, J. Y., Bernhard, P., & Ji, F. (2024). 8%—10% of algorithmic recommendations are 'bad', but… an exploratory risk-utility meta-analysis and its regulatory implications, ***International Journal of Information Management***.

Martin Hilbert, PhD, Dr; Professor, University of California, Davis, Dpt. of Communication. hilbert@ucdavis.edu

Arti Thakur, PhD candidate, Dpt. of Communication, University of California, Davis.

Pablo M. Flores, PhD candidate, Dpt. of Communication, University of California, Davis.

Xiaoya Zhang, PhD; Assistant Professor, Dpt. Family Youth and Community Sciences, University of Florida.

Jee Young Bhan; Dpt. of Sociology, University of California, Davis.

Patrick Bernhard; Dpt. of Political Science, University of California, Davis.

Feng Ji, PhD; Assistant Professor, Dpt. of Applied Psychology and Human Development, University of Toronto.

## ABSTRACT

We conducted a quantitatively coarse-grained, but wide-ranging evaluation of the frequency recommender algorithms provide 'good' and 'bad' recommendations, with a focus on the latter. We found 151 algorithmic audits from 33 studies that report fitting risk-utility statistics from YouTube, Google Search, Twitter, Facebook, TikTok, Amazon, and others. Our findings indicate that roughly 8% - 10% of algorithmic recommendations are 'bad', while about a quarter actively protect users from self-induced harm ('do good'). This average is remarkably consistent across the audits, irrespective of the platform nor on the kind of risk (bias/ discrimination, mental health and child harm, misinformation, or political extremism). Algorithmic audits find negative feedback loops that can ensnare users into spirals of 'bad' recommendations (or being 'dragged down the rabbit hole'), but also highlight an even larger likelihood of positive spirals of 'good recommendations'. While our analysis refrains from any judgment of the causal consequences and severity of risks, the detected levels surpass those associated with many other consumer products. They are comparable to the risk levels of generic food defects monitored by public authorities such as the FDA or FSIS in the United States. Consequently, our findings inform the ongoing discussion regarding regulatory oversight of the potential risks posed by recommender algorithms.

Keywords: recommender algorithms, algorithmic auditing, machine behavior, meta-analysis, digital harms.

## 1. Introduction

When Facebook whistleblower Frances Haugen disclosed internal company research with the headline "One in five teens say that Instagram makes them feel worse about themselves", Meta's Vice President of Research promptly published a response (Raychoudhury, 2021) that explained that "Although this headline emphasizes certain negative reported effects, it could have been written to note the positive or neutral effect of Instagram on users" (Facebook, 2021, p. 21). The social media platform had a point, as the same graph shows that 37 % of teens say Instagram makes them feel better about themselves. How much of algorithmic recommendations are 'good' and 'bad'?

We accept some technological risks, like the 1% lifetime death risk from cars. When risks, like the 13% lung cancer rate in smokers, become significant, we demand providers reduce them, add warnings, or limit access. This meta-analysis aggregates evidence from algorithmic audits to quantify how frequently our ubiquitous recommender algorithms pose a measurable risk.

After three decades of exponential expansion of the digital paradigm, by now, the voices pointing to its (expectedly inevitable) downsides have become louder (Lanier, 2018; Russell, 2019; Yudkowsky, 2022). One of the more concrete downsides is digitally based manipulations of the free will of human minds, manifesting itself in algorithmically induced addictions, mental health issues, misinformation, political polarization, e-commerce manipulations, online intolerance and aggression, and other human struggles (Allcott et al., 2020; Braghieri et al., 2022; Dhir et al., 2018; Nodder, 2013; Orlowski, 2020; Parr, 2015; Uzogara, 2023).

Balancing the benefits and risks of new technologies is not unique to the digital age. Realizing these trade-offs is key to understanding that technology is never inherently 'good' or 'bad' but is socially constructed (Berger & Luckmann, 1967; Kranzberg, 1986; Pinch & Bijker, 1984; Williams & Edge, 1996). As "we shape our tools and thereafter they shape us" (Culkin, 1967, p. 70), we accept risks. For instance, instead of creating slow, tank-like cars to ensure safety of the passanger, we opt for current designs despite 1.35 million annual road deaths worldwide (CDC, 2023a). Since the 1800s, when society was considering harms from stagecoach wagons, modern law has sought the right balance between technology's risks and utilities (Bergman, 2023), which let to the judicial adoption of product liability laws (Traynor, 1964). When technology features overly tilt the risk-utility balance, society reshapes them. Car manufacturers face liability for faulty airbags (Consumer Reports, 2023) or ignition switches (Stempel, 2020). Similarly, gauging algorithms' effects is vital to balance the risks they pose (Bergman, 2023).

In this study, we evaluate if the current literature can provide general estimates regarding the risks (primary focus) and benefits (secondary) of a particular digital technology: recommender algorithms. We chose this technology for four reasons: First, their societal impact, as they drive many contemporary business models (Bennett & Lanning, 2007; Davidson et al., 2010; Resnick & Varian, 1997; Ricci et al., 2011). Second, they are polemic, as they have been blamed for creating harms like filter bubbles, mental health issues, and even terrorist attack liabilities (Alfonsi, 2022; Bakshy et al., 2015; Dwivedi et al., 2021; Fletcher & Nielsen, 2018; Orlowski, 2020; Pariser, 2011; SCOTUS blog, 2023). Third, their ongoing evolution due to AI advancements, notably large language models, make it essential to take inventory of the past (Christiano et al., 2017; Ganguli et al., 2022; He et al., 2017; Vaswani et al., 2017). Last but not least, we are pragmatic. Any review requires a critical mass of existing studies, which exist for the case of recommender algorithms even in the form of new subdisciplines, such as algorithmic biases and discrimination (Caliskan et al., 2017; Hajian et al., 2016; Hannak et al., 2014), the science of fake news (Lazer et al., 2018; Vosoughi et al., 2018), and, on the methodological side, the flourishing fields of algorithmic auditing (Raji et al., 2020; Sandvig et al., 2014) and machine behavior in general (Diakopoulos, 2015; Hilbert et al., 2019; Rahwan et al., 2019; Rahwan & Cebrian, 2018).

## 2. Theoretical Framework: risk-utility

### *2.1. A theoretically impossible task*

From a theoretical perspective, no risk-utility assessment can possibly ever be complete (Bostrom, 2002; Kaplan & Garrick, 1981; Mill, 2008; Sunstein, 2005). Every assessment reflects the biases of its underlying assumptions and model. Just as it is challenging to predict a butterfly’s wing flap in Brazil causing a tornado in Texas (Lorenz, 1972), it is impossible to account for all second-, third- and n-order effects of risk and utility on all existing in- and outgroups over all time periods. In short, since risk-utility relies on models and “all models are wrong but some are useful” (Box, 1979, p. 202), its interpretation is inherently subjective, influenced by various factors from the modeler’s perception and morals, to value systems and cultural norms.

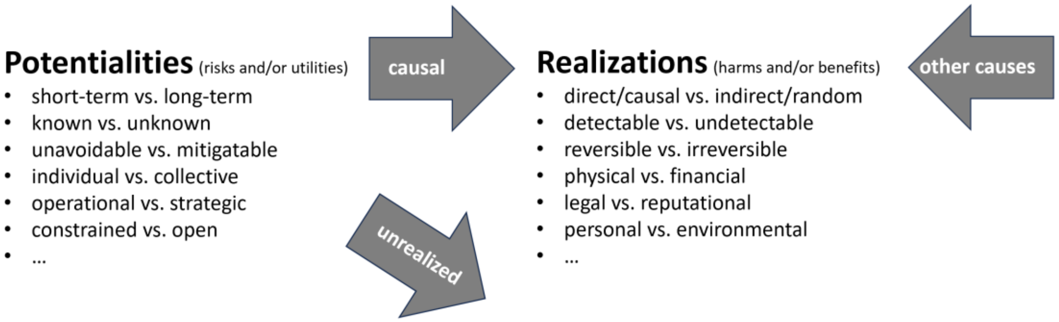
Recognizing this inherent limitation, our approach is pragmatic. We aim to encompass a broad range of perspectives and derive a general estimate that encompasses this range. We start by treating 'recommender algorithms' as a unified service to consumers, an assumption we'll validate. Our approach mirrors assessments in areas like consumer safety in food and health sectors. The perfect is the enemy of the feasible.

### *2.2. General theoretical framework*

We can differentiate between potentialities (encompassing both risks and utilities) seen as hypotheticals, and their actual outcomes (harms and benefits). Numerous aspects underpin both categories (refer to Figure 1). Pinning down causality from potentialities to outcomes is notoriously challenging, as causality can never be proven, but only rejected (Pearl, 2009) (for reasons similar to those widely accepted for hypothesis-testing (Popper, 2002)). Not every risk results in harm, and not every harm traces back to a known risk. Our study, aiming to assess the frequency of ‘good’ and ‘bad’ recommendations from algorithms, primarily addresses the potentialities of risks and utilities presented on the left-side of Figure 1.

Figure 1: Conceptual framework of potential risks/utilities and realized harms/benefits.

8%-10% of algorithms are ‘bad’…


Potentialities (risks and/or utilities)
• short-term vs. long-term
• known vs. unknown
• unavoidable vs. mitigatable
• individual vs. collective
• operational vs. strategic
• constrained vs. open
• …
causal
unrealized
Realizations (harms and/or benefits)
• direct/causal vs. indirect/random
• detectable vs. undetectable
• reversible vs. irreversible
• physical vs. financial
• legal vs. reputational
• personal vs. environmental
• …
other causes

### *2.3. What are the generally accepted levels of consumer risks?*

Most available statistics emphasize measurable harms, as seen on the right side of Figure 1. Common harm levels from everyday tools tend to be low, typically around 1% or less. Annually, there are 1.7% medically consulted injuries per 100 U.S. vehicles, with lifetime odds of dying at roughly 1% for vehicle passengers, 0.1% for motorcyclists, 0.03% for bicyclists, and less than 0.00001% for flight passengers (NSC, 2023). Nonfatal bicycle injuries requiring emergency attention are higher for 10-14-year-old males at 0.4%, compared to the 0.1% average (NSC, 2023), while bicycle helmets reduce head injury risk by 60% (Attewell et al., 2001), exemplifying ‘unrealized’ risk versus realized harm (see Figure 1). Some sectors allow for both risk (left side of Figure 1) and harm estimations (right side of Figure 1). In the Supporting Information (S.I.2. Comparison data) we present some rough ‘back-of-the-envelope’ estimates for four areas, namely food safety, consumer goods, recreational sports, and health effects of smoking. It is important to emphasize that these numbers merely provide some context for comparison. They are not the main focus of our research, nor are these areas of our expertise.

Table 1: Comparison averages of risk potential and realized harms in other areas for context (S.I.2.)

| | **Ballpark risk guestimate** | **Harmed** | **Hospitalized** | **Died** |
|---|---|---|---|---|
| Food safety: general food defects | **7%** of food samples | **15%** of U.S. citizens | **0.04%** of U.S. citizens | **0.0009%** of U.S. citizens (3,000/year) |
| Food safety: salmonella | **7.5%** of food samples | **0.4%** of U.S. citizens | **0.008%** of U.S. citizens | **0.0001%** of U.S. citizens (420/year) |
| Food safety: listeria monocytogenes | **7% - 12%** of food samples | **0.001%** of U.S. citizens | **0.0005%** of U.S. citizens | **0.0001%** of U.S. citizens (260/year) |
| Consumer goods | **0.1%** faulty goods recalled per year | **2% - 4%** of U.S. citizens | **0.6%** of U.S. citizens | |
| Sports: soccer | **3.1%** of ball possessions | **1.5%** of U.S. players | **0.04%** of U.S. players | |
| Health: smoking | **22%** of global population | Varies widely | **3% - 19%** of smokers<br>additional to general | **8.7% - 11.7%** of smokers<br>additional to general |

8%-10% of algorithms are 'bad'…

Table 1 displays socially accepted risk profiles (for sources, see S.I.2. Comparison data). In food safety, the U.S. Food and Drug Administration (FDA) maintains a Food Defect Levels Handbook, which specifies that it accepts around 7% of defect samples (mainly mold and insect-infestations). About 15% of U.S. citizens contract foodborne illnesses annually, while severe harm is much less common (some 3,000 die each year). The U.S. Food Safety and Inspection Service (FSIS) accepts 7.5% of salmonella-positive chicken carcasses and ground beef samples, with harm levels below 0.5% and around 420 U.S. deaths annually. General consumer goods have a different profile. Tracking the number of faulty products sold per year, the U.S. Dept. of Commerce only recalls 0.1%. Their accumulation and frequent use hurts 2%-4% of U.S. citizens annually. In between these extremes is the risk profile of sports. In soccer, 1 in 32 ball possessions leads to a potentially dangerous foul (3.1%), and 1.5% of U.S. players end up injured. Another extreme is the risk profile of cigarettes. While each cigarette is risky and severe harm levels are notably higher (some 10% die from lung cancer or cardiovascular disease), it is surprising to many that these trackable risks are not life-threatening to some 90% of smokers.

### *2.4. What do we know about quantifiable digital risks?*

In the realm of information systems, data shows wide variance. Studies indicate 2.6% to 3.8% of global e-commerce orders in 2021 were fraudulent (Pasquali, 2022), while digital ad security violations were between 0.14% and 0.22% (Statista, 2023a). Regarding realized harms, 4% to 8% have experienced bank fraud, with under 1% facing it more than three times (Petrosyan, 2022). Identity theft rates in the U.S. range from 0.08% to 0.6% annually (Duffin, 2023). Such variations are likely due to the field's novelty, lack of clear definitions, and no public authority charged with overseeing potential risks.

One area of much active research (which is one of the areas covered in our inventory) links social media use and health outcomes (see several literature reviews covering hundreds of studies (Frost & Rickwood, 2017; Haidt et al., 2023; Orben, 2020; Sohn et al., 2019; Yoon et al., 2019)). Most studies find a small negative association between digital tech or social media, and well-being, roughly $r \approx -0.12$ (McCrae et al., 2017; Meier & Reinecke, 2021; Orben, 2020). While small, it is within the ballpark of traditional public health concerns: the correlation of childhood lead exposure and adult IQ is 'merely' $r = -.11$ (Reuben et al., 2017), and has resulted in more than 30 years of public oversight, with current ambitions to reduce the risk even further for special interest groups like small children (FDA campaign 'Closer to Zero' (FDA, 2023a, 2023b)). From the perspective of media studies, the rather small effect is also not surprising: differential susceptibility to media effects is one of the most commonly accepted and longstanding media theories (Valkenburg & Peter, 2013). In other words, it must be expected that recommender algorithms do not affect everyone to the same degree and that the inherent diversity lowers the overall correlation. We are just learning about conditions for causal links (Nisar et al., 2019). For example, adolescent minds seem to be more susceptible (Haidt et al., 2023). In S.I.2., we offer an estimate of the 'Self-reported social media harms by adolescents', which averages at around 20%. This number would be equivalent to the numbers presented in the middle column of "harmed" users in Table 1. Only general food defects reach a similar order of magnitude of self-reported harms.

Moving on to what we know about the right column in Table 1, including death, in the extreme case, the debate about causal links is currently discussed in courts (60 Minutes, 2022; Ortutay, 2023; SMVLC, 2023). For example, the link between digital conduct and suicidal ideation is hotly debated (Memon et al., 2018; Twenge et al., 2020), as female high school students who have 'seriously considered attempting suicide' increased from 19% in 2011 to 30% in 2021 (with 13% actually having attempted suicide) (CDC, 2023b). It is important to emphasize that not all recently appearing mental health risks are directly tied to recommender algorithms. For example, it is often impossible to distangle the causal effects of the COVID-19 pandemic from the parallel rise of social media in observational data.

Deaths from online challenges, like the "Pass Out Game" or "Blackout Challenge" where participants asphyxiate themselves, are often mislabeled as suicides or accidents, which is the reason why official numbers of participants hanging themselves on dog leashes or suffocating from locking belts are likely underestimates. As early as 2008, the U.S. Centre for Disease Control and Prevention (CDC) already cites 82 dead youth victims (Glasper, 2023) and The popularity of the 'online challenge' rose thereafter. In 2016, a non-profit in honor of one victim claimed to have identified some 1 million search results for videos on 'How to play the Pass Out Game' (EricsCause, 2023), various with several hundred thousand views. The organization offers an unverified map of victims that lists the names and locations of the deaths of 981 individuals (mostly adolescents) since 2005 (the birth year of YouTube), with 685 registries from the U.S. alone. Other deadly social media challenges include the 'tide pod challenge' (consume laundry detergent), 'skull breaker challenge' (two people tripping a third), 'fire challenge' (set yourself on fire), 'outlet challenge' (put penny in outlet), 'Benadryl challenge' (inducing hallucinations by overdosing), among others. While the total number of victims from such challenges is not known (because often misclassified as suicides or accidents), we know that it compares with 260 deaths from listeria, 420 deaths from salmonella, and a total of 3,037 from general foodborne illness in the U.S. last year (see S.I.2. Comparison data). In contrast to severe social media harms, these harms are officially monitored.

While mental health is rather straightforward to diagnose, there are other possible risk that are more difficult to measure. For example, social media and polarization recently filled an entire special issue in the journal Science (Uzogara, 2023). Such measurement challenge applies not only to the quanitification of risk, but also to the evaluation of utilities. A recent review of some 500 articles emphasizes that threats to democracy need to be balanced with the utility of increased participation (Lorenz-Spreen et al., 2023).

With all of this in mind, our inventory focuses on the left-side of Table 1, quantifying (what in Table 1 we refer to as the) "ballpark risk guestimate". We only focus on those stemming from recommender algorithms, but within that rubric, we have the declared ambition of obtaining a ballpark measure of a wide range of recommendation algorithms. Our final classification aims at harmonizing risks as diverse as those posed by unsafe child content, job-market discrimination, misinformation and conspiracies, mental health issues, and political radicalization. Therefore, we purposefully use the exaggeratedly trivial classification terms of 'good' and 'bad'. This simplistically naïve language explicitly serves as a blatant reminder to avoid a false sense of precision in our subjective translation from the research findings into our aggregate framework. As explained more in detail for each case in 'S.I.1. Data repository', we understand 'bad' not to be a recommendation the user does not like, but rather as potentially harmful content. In short, we do not quantify the realized harm to people but aim to obtain a very generic ballpark number of risks and utilities provided by algorithmic recommender systems. In this study, we also do not make any claims for causal links between risk and harm, or between utility and benefit. We had the ambition to move toward more quantitative and cross-platform comparisons and employ –what Jeyaraj and Dwivedi (2020) call– an exploratory meta-analysis, moving beyond the single existing review study we had found, which served as an inspiration (Yesilada & Lewandowsky, 2022).

## 3. Method

As our analysis aims at bridging measures from computer science, information science, medicine, and the psychological and social sciences, we prioritized expected values and simple conditional probabilities over higher-order meta-analytic statistics (Higgins et al., 2019; Petticrew & Roberts, 2008; Uman, 2011). We still achieve the meta-analytic goal of systematically synthesizing independent studies to calculate an overall effect (Egger & Smith, 1997; Shorten & Shorten, 2013). Table 2 presents the simple framework that conditions algorithmic recommendation output on different kinds of input. While all included studies (N = 151) report the percentage of ‘bad’ recommendations (first column: X%, Y%, or Z%), we only obtain data on ‘good’ recommendations for 62 studies (see Table 3). This means that we can distinguish between ‘bad’ and ‘not bad’ recommendations for all 151 audits (which is what we will do for most of our analyses), and distinguish between ‘good’, ‘other/neutral’, and ‘bad’ for a subgroup of studies (see section ‘Recommending utility content’).

Table 2: Risk-utility framework

| | | **output** | | | |
|---|---|---|---|---|---|
| | | ‘bad’ | ‘not bad’ | | |
| | | | ‘good’ | ‘other/neutral’ | **total** |
| **input** | ‘bad’ | $X_b$ % | $X_g$ % | 100 - $X_b$ - $X_g$ % | **100 %** |
| | ‘other/neutral’ | $Y_b$ % | $Y_g$ % | 100 - $Y_b$ - $Y_g$ % | **100 %** |
| | ‘good’ | $Z_b$ % | $Z_g$ % | 100 - $Z_b$ - $Z_g$ % | **100 %** |
| | **total** | **$T_b$ %** | **$T_g$ %** | **100 - $T_b$ - $T_g$ %** | **100 %** |

### *3.1. Searching the Literature*

We began with searches using the databases of Google Scholar, arXiv, SSRN, PubMed, EBSCO-ERIC, ProQuest, and PsycINFO, using any combination of keywords like recommender algorithms/ machines/ engines, algorithmic auditing, algorithmic bias, misinformation/ conspiracy/ extremism/ hate, mental health, polarization, children/ kids, Facebook, Instagram, Netflix, Snapchat, TikTok, Twitter, YouTube, among others, between 01/15/2023 and 03/23/2023. We identified 69 articles and started snowball method searches based on the obtained references, using the AI-tools of elicit.org, researchrabbit.ai, connectedpapers.com, and Google Scholar. Our biggest challenge was to find studies that allowed us to calculate the input-output transition probabilities of Table 2. We also excluded some studies for reasons of lack of rigor or transparency, but kept five studies that were not peer-reviewed after we did a peer-review ourselves. We were left with 151 audits from 33 articles (see S.I.1. Data repository). Except for two studies (from 2012 and 2016), all were published within the last five years (2018 – 2023).

### *3.2. Coding the Literature*

We are transparent about the subjective judgments we used to translate the reported statistics into our framework and present them in S.I.1. Data repository. We reached these judgments through collective deliberations in weekly research meetings among the seven co-authors. In addition to the transition frequencies to fill Table 2, we characterize the 151 audits by the underlying platform/company, the type of content, and the modalities of recommendation (see Table 3 and S.I.1. Data repository).

Table 3: Descriptive variables of recommendations

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Recommendation output** | **bad** | **not bad (other)** | **good** | | | | |
| | 151 | 151 (89) | (62) | | | | |
| **Proceeding input** | **Bad** | **other/ neutral** | **good** | | | | |
| | 50 | 80 | 21 | | | | |
| **Platform / company** | **YouTube** | **Search Engines** | **Twitter** | **Facebook** | **TikTok** | **Amazon** | **Instagram** |
| | 86 | 21 | 15 | 13 | 13 | 2 | 1 |
| **Type of content** | **mis-information** | **political radicalization** | **child harm** | **bias/ discrimination** | **mental health** | | |
| | 67 | 38 | 20 | 14 | 12 | | |
| **Modalities of recommendation** | **search results** | **autoplay** | **API graphs** | **newsfeed** | **side panels** | **channels** | **homepage** |
| | 44 | 26 | 29 | 19 | 18 | 8 | 7 |

*of the 151 ‘other/neutral’ audits, 84 do not distinguish between ‘good’ and ‘other/neutral’ recommendations, while 62 do.

## 4. Results

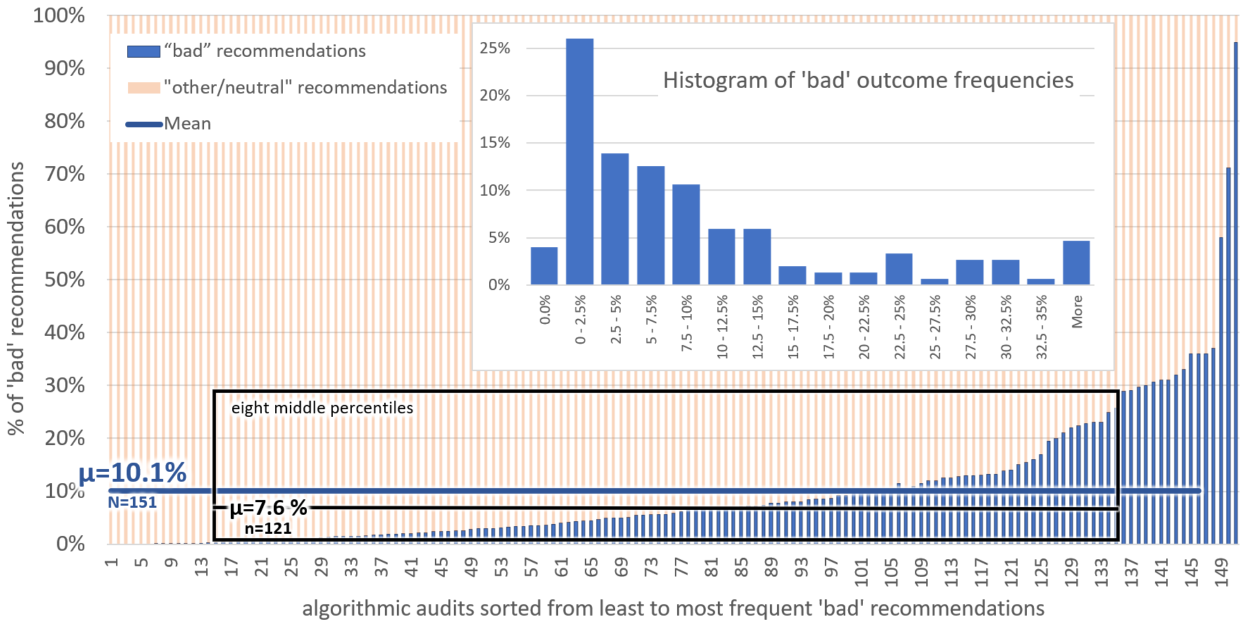
In general, our analysis conditions the obtained recommendation frequencies (see Table 2), on the descriptive characteristics of the application (see Table 3).

### *4.1. Recommending ‘bad’ content*

The histogram in Figure 2 shows that most of the audited recommender algorithms produce between 0% and 2.5% of ‘bad’ recommendations, and over two-thirds of audits (68%) have a rate of less than 10 %. The resulting skewed and asymmetric distribution of ‘bad’ recommendations centers around a mean value of M = 10.10 % (SD = 13.2 %; Median = 5.9 %; skewness 3.0; kurtosis 13.3) (see Figure 2). Equally excluding extremes on both ends by eliminating the top- and bottom-percentiles reduces the mean to 7.6 % (SD = 6.4%) (see Figure 2), which confirms that some ‘bad’ outliers weigh more heavily. Excluding the three rightmost outliers reduces the mean to 8.9 % (SD = 9.5 %), and moves the resulting distribution within the accepted levels (Hair et al., 2009) of normality (skewness 1.4; kurtosis 1.2).

8%-10% of algorithms are ‘bad’…

Figure 2: Distribution of ‘bad’ recommendations.



Of course, any concern about realized harm would argue against excluding outliers. However, both the statistical shape of the empirical distribution and substantive reasons linked to the particularities of these three cases[1] suggests that it would be helpful to exclude them going forward. When in doubt, we will continue to check for their influence on our results but exclude them otherwise.

### *4.2. By platform*

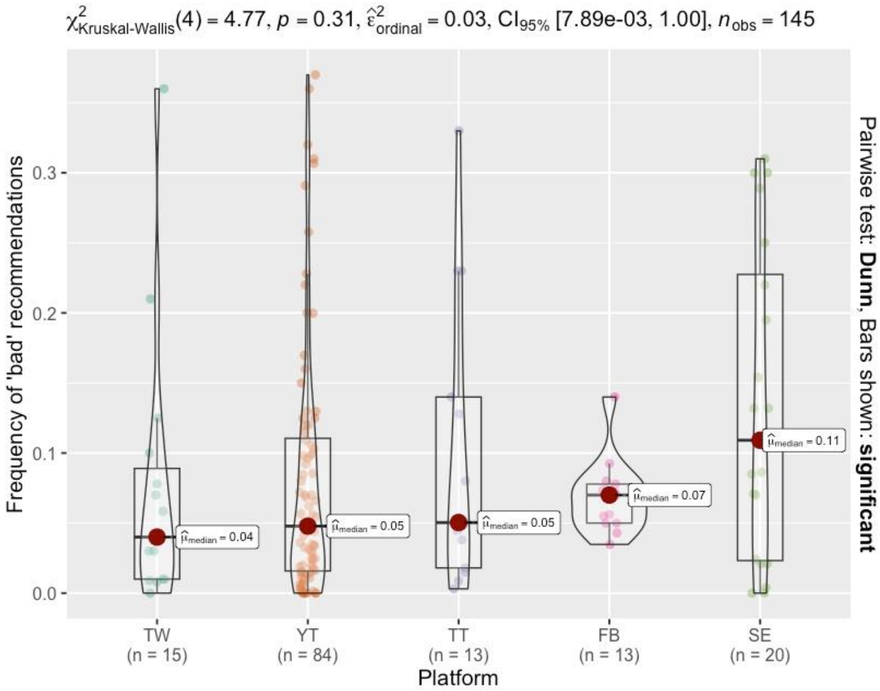
For an analysis among platforms, we exclude Amazon and Instagram due to small samples (see Table 3). The Kruskal-Wallis test (n = 140) indicated no significant difference among the frequency of ‘bad’ recommendations among different platforms, $\chi^2(4) = 4.77, p = 0.31, \varepsilon^2 = 0.03$.[2] None of the post hoc tests is significant (Dunn with Bonferroni correction, 0.65 ≤ p ≤ 1.0) (see Figure 3). This serves as a post hoc justification for our methodological decision to analyze recommender engines from different companies together, as a single concept/ service/ product.

[1] The rightmost outlier (94.9% of 'bad' recommendations) comes from a very early 2016 YouTube Kids study (Kaushal et al., 2016), measuring recommendations of content deemed unsafe for kids when starting at content inappropriate for kids (having "graphic nudity or abusive/inappropriate dialogues", p.1). YouTube Kids was just founded in 2015. The second outlier (72%) measures the promotion of conspiracy content by the Russian search engine Yandex after searching for conspiracy content, like 'flat earth', 'qanon' '9/11', or 'illuminati' (Urman et al., 2022). The third outlier (58%) is a YouTube study that measures the promotion of adult content when watching the popular kids' channel 'Ryan's world' (Merrer & Trédan, 2022). While this top-10 YouTube channel aims at children 2-6 years old (“Ryan’s World,” 2023), its consumer focus on unboxing content also caters to parents. On the other side, we find seven audits with 0% ‘bad’ recommendations, including two of political radicalization of alt-right fringe ideas like that of a white ethnostate (Ribeiro et al., 2020).

[2] While Kruskal-Wallis is quite robust against violations of the normality assumption, we supplemented the result by running Welch-ANOVA ($F(4, 37.7) = 1.7, p = .17,\ \omega^2 = 0.06$). None of the post hoc tests is significant (Games-Howell, p = 1.0).

8%-10% of algorithms are ‘bad’…

Figure 3: Frequency of ‘bad’ recommendations per platform/company (n = 145). TW (Twitter), TT (TikTok), YT (YouTube), FB (Facebook), and SE (Search Engines).



We still note the higher mean for Search Engines (SE, M=13%), versus Twitter (TW, M=7.6%), which might also be due to the nature of search engine audits, which, often, specifically search for ‘bad’ content (like conspiracies or misinformation).

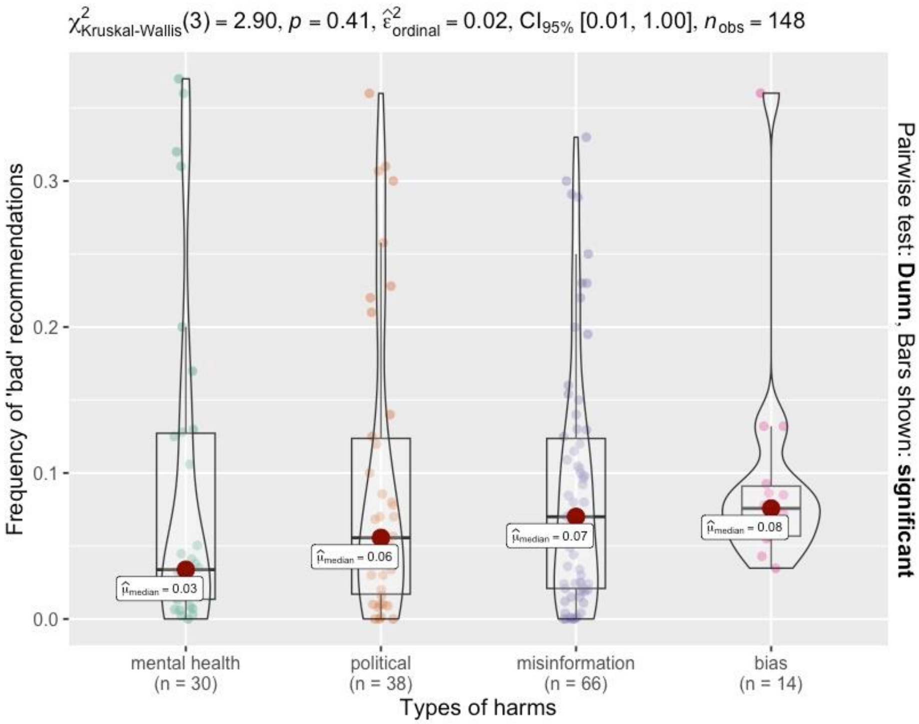
### *4.3. By content types of risk*

Our inventory joined a notable variety of quite different risks, which warrants the question of whether we are mixing apples with oranges that should be treated separately. A Kruskal-Wallis test among the frequency of different types of 'bad' recommendations is not significant between audits that looked at ‘mental health’ (including kids’ studies), ‘political content’, ‘misinformation’, and ‘biases’ (discrimination), $[\chi^2(3) = 2.9, p = 0.4, \varepsilon^2 = 0.02]$. None of the post hoc tests is significant (Dunn with Bonferroni correction , 0.5 ≤ p ≤ 1.0) (Figure 4).[3] Of course, as one of our peer-reviewers aptly highlighted, this is like joining the health risks arising from rotten oranges and apples with worms. It is our declared goals to do this for recommender algorithms. Hence, from our overarching risk-assessment perspective, it serves as a post hoc justification for our methodological decision to analyze recommender engines as a single defining group in terms of their potential risk rate.

[3] Welch-ANOVA: [F(3, 47.86) = 0.22, p = .88; $\omega_p{}^2$= 0]. None of the post hoc tests is significant (Games-Howell).

8%-10% of algorithms are ‘bad’…

Figure 4: Frequency of ‘bad’ recommendations per type of risk (n = 148).

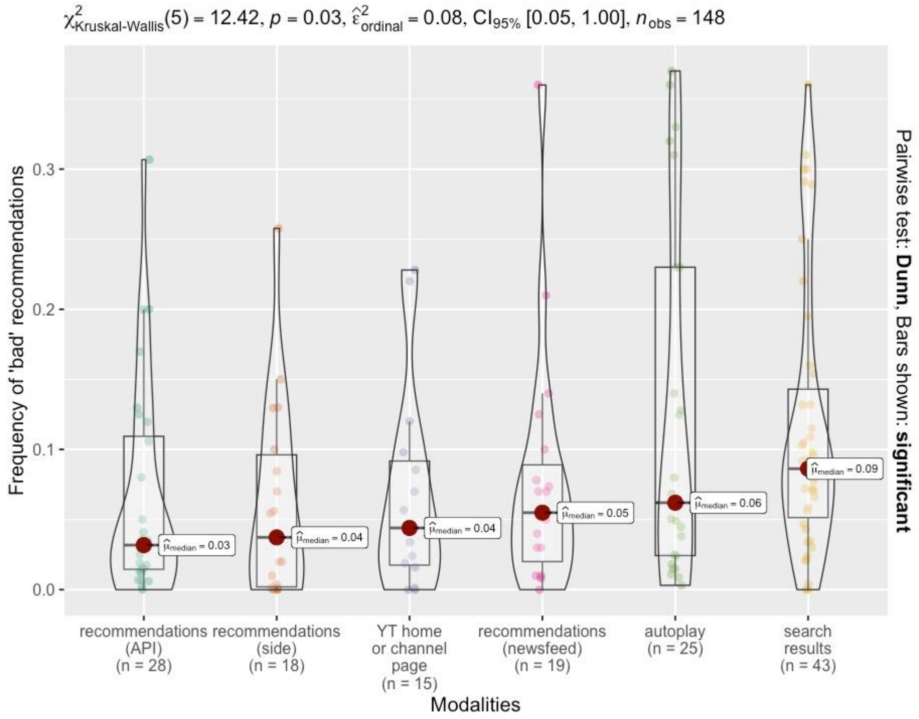


### 4.4. By modalities of recommendations

It is important to note that the output of recommender engines reaches the consumer in different ways. Some are more or less subtle (a side banner versus a homepage), while others require more or less proactive interaction (a search result versus autoplay) (Figure 5). We performed a Kruskal-Wallis test to compare different modes of ‘bad’ recommendation, followed by a Dunn's post-hoc test for pairwise comparisons. The Kruskal-Wallis test revealed a significant difference among the groups ($\chi^2(5) = 12.4, p = 0.03, \varepsilon^2 = 0.08$).

8%-10% of algorithms are ‘bad’…

Figure 5: Frequency of ‘bad’ recommendations per modality application of algorithms (n = 148).

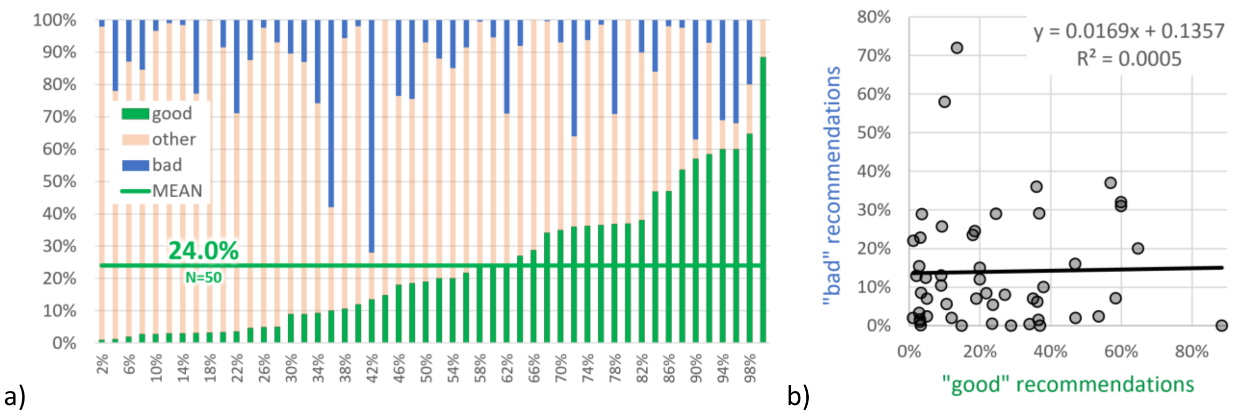


While Dunn's post-hoc test, with Bonferroni correction for multiple comparisons, did not find significant differences (Dunn, p<.001), it is interesting that the two extremes are recommendations (API) (M=6.5%) and search results (M=11.5%). One YouTube audit found that the “YouTube Data API results are similar to those of the non-logged-in profile with no watch history (using a browser). This indicates that recommendations returned using the API are not subject to personalization” (Papadamou et al., 2022, p. 724). On the other extreme, users provide proactive input to guide the personalized result, steering each recommendation with custom-made input. In several of the audits, researchers explicitly searched for risky content (such as "flat earth" or "vaccines and autism"), which is likely to contribute to the above-average ‘bad’ frequency of search results.

### *4.5. Recommending utility content*

While most studies focus on risk, i.e. ‘bad’ content, we agree that it “may not be possible to build a system that avoids ‘bad’ without also defining ‘good,’ because of the necessity of specifying objectives” (Stray et al., 2022, p. 4). The foregoing analysis merged all ‘non-bad’ recommendations into a single ‘other/neural’ group. A smaller subset of audits (including the Yandex and Ryan’s world outliers) also distinguished between ‘good’ (or ‘utility’) recommendations in contrast to ‘bad’ (or ‘risk’), which allows us to create three groups. As shown in Figure 6a, the average rate of ‘good’ recommendations is notably higher and passes as normally distributed (M = 24; SD = 20.7%; skewness 0.97; kurtosis 0.46). The majority of audits (54%) produce between 9% and 38% of ‘good’ recommendations. The rightmost outlier (89% of 'good' recommendations) comes from a very early 2012 Google search study that evaluates websites providing correct answers to influenza vaccine searches (Betsch & Wicker, 2012). The second outlier (65%) is also a vaccine study focusing on vaccine-related YouTube content (Abul-Fottouh et al., 2020). The third and fourth positive outliers (60%) measure the promotion of kids' content when watching two popular kids' channels on YouTube (Merrer & Trédan, 2022). On the other extreme of low ‘good’ recommendations, we find studies of misinformation on Google Search (Zade et al., 2022), political radicalization on YouTube (Ribeiro et al., 2020), and vaccine misinformation on Amazon.com (Juneja & Mitra, 2021).

Figure 6: Distribution of ‘good’ and ‘bad’ recommendations (n = 50, including outliers), (a) bar graph; (b) scatter plot.



It is interesting to note that there is no correlation between the amount of ‘good’ and ‘bad’ recommendations (Figure 6b). This finding seems to suggest that there might not be any inherent hindrance to designing recommender algorithms that minimize risks and maximize utilities (Bergman, 2023; Hylton, 2012).

### *4.6. By proceeding seed input*

Following the different rows of Table 2, we now analyze if the frequency of resulting ‘good’ or ‘bad’ output recommendations depends on the ‘good’ or ‘bad’ nature of the proceeding seed input.

We start by looking at different kinds of input leading to 'bad' output recommendations (Figure 7a). None of our tests finds a significant difference between 'bad' input (M=10.5%, SD=9.6%; median = 8.5%), 'other/neutral' input (M=8.1%, SD=8.3%; median = 5.6%), and 'good' input (M=9.5%, SD=13.1%; median = 2.2%).[4] Despite this statistical finding, we note that the median of ‘good’ input is notably lower than

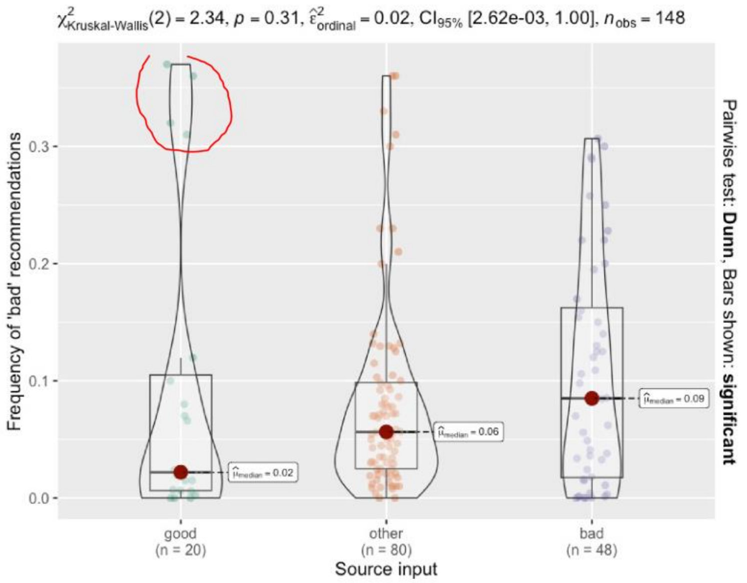
[4] Running different tests to cover the strength and weaknesses discussed in (Tomarken & Serlin, 1986):

the rest and that there seem to be some outliers on the higher side of ‘bad’ recommendations (see Figure 7a). It turns out that all four of the potential audits marked in Figure 7a stem from the same study (Merrer & Trédan, 2022), a YouTube study that assessed the likelihood that a user ends up with content not for kids when following autoplay recommendations after starting at a kids’ channel. It can be argued that starting at a kids’ channel on general YouTube (not YouTube Kids) is not enough of a ‘good’ input to keep one away from content not made for kids.

We repeat the analysis without these four cases (N=139), and find a significant difference $[\chi^2(2) = 10.1, p < 0.001, \varepsilon^2 = 0.07]$[5] with significant post hoc effect (Dunn, p<.001) between ‘bad’ input (M = 10.5%; SD = 9.6%; median = 8.5%) and ‘good’ input (M = 3.3%; SD = 4.0%; median = 1.5%).

We do the same for ‘good’ output recommendations on the smaller dataset (n= 50) (Figure 7c). Here we find a significant difference even with including these four data points $[\chi^2(2) = 11.63, p < 0.001, \varepsilon^2 = 0.24]$ ,[6] with significant post hoc effect (Dunn with Bonferroni correction, p<.001) between ‘bad’ input, (M = 14.6%; SD = 12.4%; median = 12.4%), and ‘good’ input (M = 37.5%; SD = 15.7%; median = 36.5%). We confirm the result when excluding the four data points from the YouTube children's content study (Merrer & Trédan, 2022).[7]

Figure 7. Conditioning on proceeding input-source seed for (a) ‘bad’ recommendations (n=148), with marking of four outlier; (b) without outliers (n=144); (c) ‘good’ recommendations (N=50).



(a)

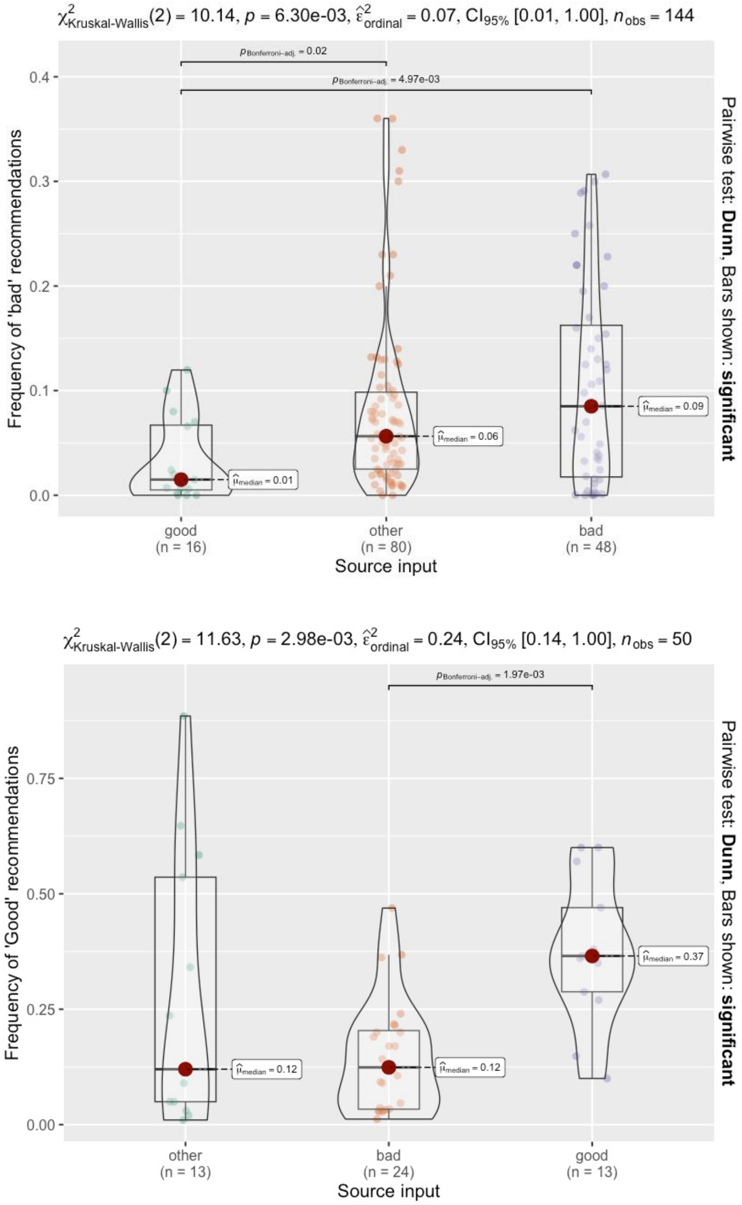
Kruskal-Wallis: $[\chi^2(2) = 2.3, p = 0.31, \varepsilon^2 = 0.02]$, without a single significant post hoc effect (Dunn with Bonferroni correction). Welch: F(2, 45.66) = 1.3, p = .37. Brown-Forsythe: F(2, 46.6) =.93, p = .40. None of the post hoc tests with Games-Howell is significant, .7 ≤ p ≤ 1.0.

[5] Welch: F(2, 60.5) = 10.5, p < .001. Brown-Forsythe: F(2, 94.98) = 6.51, p = .02.

[6] Welch: [F(2, 21.33) = 10.34, p < 0.001]. Brown-Forsythe: F(2, 21.0) = 5.9, p = .009.

[7] Kruskal-Wallis $[\chi^2(2) = 12.7, p < 0.001, \varepsilon^2 = 0.027]$ ,[7] with significant post hoc effect (Dunn with Bonferroni correction, p<.001) between ‘bad’ input, and ‘good’ input.

8%-10% of algorithms are ‘bad’…



(b)

(c)

## 5. Discussion

Within the larger context of our theoretical development (Figure 1) explored in Sections 2.2 and 2.3 (Table 1), our main finding (Figure 2) suggests that recommender algorithms have a risk profile akin to food safety. The literature tells us that self-reported social media harms for adolescents, for example, are around 20% (see S.I.2. ‘Self-reported social media harms by adolescents’), which is expected to be higher than the general population average. The identified harm level of general food defects is in the 15% ballpark area (see Table 1). While we lack a clear accounting for the most severe realized harms experienced by social media victims, anecdotal evidence fuels the dreadful suspicion that physical harms could affect hundreds (or even thousands) of (often yet unidentified) users (including deaths, such as stemming from social media challenges, see discussion in section ‘What do we know about quantifiable digital risks?’). The ballpark area of realized harms for food safety, salmonella, and listeria is between 260 and 3,000 deaths per year (Table 1).

Our study aimed to fill a gap in the literature that referred to the ‘left-side’ of Figure 1, which refers to potential risks. We can now add that, also here, recommender algorithms end up in a similar ballpark risk profile as food safety. Health audits find that between 7% and 12% of food samples are defective. This does not mean they necessarily harm anyone. Being a mere potentiality, audits detect samples to be ‘bad’, i.e. moldy, or insect- or salmonella infected, etc. Our Figure 2 finds that algorithmic audits find the general risks of ‘bad’ recommendations to be between 8% and 10%. Also here, this does not imply that these recommendations do any harm, but auditors detected that they are potentially risky. We discuss the practical implications in section 5.2 below.

### *5.1. Contributions to Research*

Our concrete findings highlight the consistent level of risk from recommender algorithms across different platforms (Figure 3) and risk types (Figure 4). This consistency supports the idea of treating, and perhaps regulating, these algorithms as a unified category of consumer products. However, this overarching risk consistency masks the qualitative differences in actual harms (compare Figure 1). It is tricky to compare the damage from misleading vaccine information (Betsch & Wicker, 2012; Juneja & Mitra, 2021; Tomlein et al., 2021) with that from race and gender biases in job searches (Ali et al., 2019; L. Chen et al., 2018), or to contrast the promotion of ethnic intolerance (Albadi et al., 2022) with the long-term harms of exposing hundreds of thousands of toddlers to violent and sexual content (Papadamou et al., 2020).

Regarding how recommender algorithms are employed (Figure 5), we only find significant differences between recommendations drawn from impersonalized API data and tailor-made specific searches. While API data is not subject to personalization (Papadamou et al., 2022), audits of pinpointed searches explicitly search for risky content, so the difference is not very surprising. While the autoplay option has the largest mean (M=12.2%, SD=12.6%), more than twice as large as recommendations derived from API access (M=5.4%, SD=7.2%), the difference is not significant. The autoplay option's high average is influenced by four audits from a YouTube study examining transitions from child to non-child content (Merrer & Trédan, 2022). Excluding these decreases the autoplay average to M=8.1 (SD=8.8%). This underscores that our small sample size constrains our findings, and more studies are needed.

Another finding is the clear evidence that recommender algorithms do more ‘good’ than ‘bad’ (compare Figures 1 and 5). Their benefits drive their popularity, a fact often overshadowed by concerns about risks (Alfonsi, 2022; Orlowski, 2020; SCOTUS blog, 2023; The New York Times, 2020). This finding suggests to focus on these positives and measure them more consistently, as many risk-audits overlook them. There is much societal utility in the systematic employment of recommender algorithms that are less biased than a human judge (Hajian et al., 2016), more balanced than existing political discourse (Bakshy et al., 2015), or offer more accurate medical recommendations than trained experts (Esteva et al., 2017).

8%-10% of algorithms are 'bad'…

An intensified focus on positive content is also justified when comparing the good/bad nature of the input (Figure 7). While the hotly debated 'rabbit hole hypothesis' (The New York Times, 2020) suggests users get trapped in a negative recommendation cycle, our data shows that –after removing an outlier– 'bad' to 'bad' transitions occur three times more than 'good' to 'bad' ones (M=10.9% versus M=3.3%). Note that the outlier we removed focused on kids content, which suggests that children might be particularly in danger of being algorithmically guided from 'good' content, to 'bad' content. Switching the focus from 'bad' recommendations to 'good' recommendations, we found clear-cut evidence of an upward spiral that locks users algorithmically into a cycle of 'good' recommendations.  When starting at 'good' input content, there is an almost 40% chance to receive 'good' recommendations versus only 15% when starting as a 'bad' content seed. This suggests that the risk of a negative downward spiral that drags users into the rabbit hole exists but that –at the same time—an even larger positive upward spiral also drags users out who brought themselves down, interjecting more frequent upside content.

### *5.2. Implications for Practice*

We draw two main practical implications from our findings: (1) we find a level of potential risk that –in other areas of modern life—is overseen by a public regulatory agency, hence we suggest creating one; (2) we find significant variations in risks levels, which suggests that algorithm providers can likely tweak them to minimize risk and enhance utility. We elaborate on both points.

In an influential article from 2017, the attorney-adviser of the Office of Legal Counsel of the U.S. Department of Justice, Andrew Tutt, calls for "An FDA for Algorithms". The article "proposes that certain classes of new algorithms should not be permitted distributed or sold without approval from a government agency designed along the FDA. This 'FDA for Algorithms' would approve certain complex and algorithms when it could be shown that they would be safe and effective for their use and that satisfactory measures would be taken to prevent" (Tutt, 2017, p. 83). In agreement with common FDA practice, the acceptable rate of failure does not need to be zero. As discussed in more detail in 'S.I.2. Comparison data', the FDA's Food Defect Levels Handbook accepts, for example, a median 7% of samples containing mold, insect-infestations and other defects (see also Table 1).

Having a publicly sponsored, competitively neutral space, such agencies usually also serve as a centralized body to coordinate expertise that helps to develop guidance, standards, and best-practices in partnership with industry, striking a balance between innovation and safety. Such agencies also elevate awareness of algorithmic influence (Shin et al., 2022). It will naturally create a topical clearing house, as understanding algorithms "requires technical knowledge of recommender design and operation, and also critically depends on insights from diverse fields including social science, ethics, economics, psychology, policy and law." (Stray et al., 2022, p. 1). Such space can facilitate the long overdue discussion about the definition of acceptable levels of risk when operating algorithms.

Often, any call for more regulatory oversight is shut down by the argument that modern algorithms are 'black boxes' (Guidotti et al., 2018; Pasquale, 2015) and that we first need to be able to explain how they work before anything can be done. While the focus on explainable algorithms is laudable and intellectually stimulating, most of the other regulated risks of modern life do not require such high stakes. Health regulators do not demand to reveal a soda's recipe to link it to obesity. Famously, Coca

8%-10% of algorithms are 'bad'…

Cola's secret formula, guarded for over 135 years (World of Coca Cola, 2023), has not stopped authorities from studying its health effects. Drug administrations do not necessarily require revealing patented medicine recipes to execute randomized controls experiments that study its health effects. All of the reviewed studies take a similar approach, applied to machine behavior (Diakopoulos, 2015; Hilbert et al., 2019; Rahwan et al., 2019; Rahwan & Cebrian, 2018) through algorithmic auditing (Raji et al., 2020; Sandvig et al., 2014). Evaluating risks and harms on humans does not require understanding the black box that produces them.

This leads to the question if it would be possible to reduce the risk without losing utility. Our findings suggest that it would be possible: we found a wide diversity in 'bad' and 'good' content and their combination. In fact, we did not find any correlation between the level of 'good' and 'bad' recommendations (Figure 6b), which suggests that "alternative designs exist that offer equivalent utility and less risk" (Hylton, 2012, p. 2458). For example, we found search engine algorithms that produced some 25% of 'bad' results and only 3% of 'good' results (Makhortykh et al., 2022), and other best practices that produced over 50% 'good' results and only 2% of 'bad' results (Betsch & Wicker, 2012; Urman et al., 2022). At the very least, this suggests that higher utility does not inherently mean greater risk.

One practical illustration of this last point is exemplified by the increasing focus on value-sensitive design (Friedman et al., 2002), such as RLHF (Reinforcement Learning from Human Feedback) (Christiano et al., 2017; Ustalov et al., 2022). For example, before releasing ChatGPT-4.0, OpenAI 'aligned' it for eight months, between August 2022 and March 2023, to reduce potential risks. Its final 'alignment' report explained that "GPT-4 produces toxic generations 0.73% of the time while GPT-3.5 produces toxic generation 6.48% of the time" (OpenAI, 2023, pp. 22–24). The private company reduced the risk by almost a factor of 10, just by adding some additional value-sensitive design. Some studies suggest that human interaction has reduced performance and accuracy (L. Chen et al., 2023), which would also suggest reduced utility. Of course, no solution will be easy and will depend on values, that have to be proactively created, such as exemplified by the ongoing discussions about content moderation (Gillespie, 2018, 2020; Myers West, 2018). These emerging discussions about the practical implications of risks and utilities of intelligent algorithms show the urgent need for more discussions in academia, within the private sector, and through the existing channels of the executive, judicial and legislative branches.

### 5.3. Limitations and Future Research

We see several limitations of our study, referring to our sample, to the coarse-grained nature of our review, the lack of fine-grained and accumulative effects, and to causality.

Our data can be expected to be biased toward higher levels of risk than a truly random sample, due to studies that focus on risk-prone communities by design (Papadamou et al., 2021). It is redundant to mention that our review is certainly also not exhaustive. We anticipate finding more studies as soon as we finish the analysis. We also expect new fields of harm to be discovered that we are just starting to understand (e.g. (Akter et al., 2021)). Despite our extensive approach, our small sample sizes mean some results may hinge on specific studies, as seen with Figure 7. Such results warrant caution, and further research will be crucial for understanding finer details.

Our main variable of 'good' and 'bad' recommendations is as 'good' and 'bad' as this purpusfully simplistic distinction. This broad brush misses many important aspects. Often good is a generic term for no harm found, but most studies only seek one specific harm, which never guarantees the absence of potential other harms. Additionally, 'bad' as a category simplifies important distinctions of 'badness', as gender-disrimination, child-harms, or misinformation are certainly different in nature. While our aim for a summary figure led to less precision, opting for a broad variety of studies was intentional. Despite excluding some detailed higher-order measures of effect sizes and confidence intervals from audits, the consistency across our classifications (see Table 3) validates our method for this initial meta-analysis. Being aware that our exploration is rather a framework in search of evidence, we hope it is useful to generate more tailored and comparable studies on algorithmic risks and benefits.

The measure we summarize, typically reports a snapshot of individual recommendations at a given time, not cumulative effects. For context, while every single cigarette is 'bad' (containing nearly 70 carcinogens), one cigarette alone is unlikely to cause cancer. Something similar can be expected to apply to 'bad' recommendations. Our review suggests that one 'bad' recommendation daily can be expected when consuming 10-13 algorithmic recommendations a day (8-10% of recommendations). Given 150-minute average daily social media use in 2022 (Statista, 2023b), users viewing 1-minute autoplay videos (like TikTok, YouTube shorts or Instagram reels) might encounter 10-15 'bad' videos daily. Watching longer 4-minute YouTube videos, half being algorithmically chosen (Hosseinmardi et al., 2021), would result in 1-2 'bad' algorithmic recommendations per day. Taking the simple average of this coarse-grained guestimate suggests that the average user might receive some 7 'bad' video recommendations per day. Comparatively, the average smoker smokes some 6 'risky' cigarettes per day (Worldometer, 2023). We do not understand and do not make any claims about the accumulative effects of such risk, but argue that it might be worth considering accumulative effects in future studies. It might turn out that cumulative effects play a cricual role in the conversion of potential risks to realized harms, as they do in other areas of recurrent consumption of risky products.

In the same sense, we explicitly refrain from making claims of causality in this analysis (we stay on the 'left-side' of Figure 1). The complexity of harmful algorithmic recommendations' causal effects still requires more largescale systematic statistical research, as well as in-depth analysis of individual cases of potential victims. Media studies' theory tells us that we can expect that media effects are highly differentiable between different humans with different usage patterns (Valkenburg & Peter, 2013). Similarly to the fact that over 80% of smokers do not experience the severe harms from smoking presented in Table 1, we know that the other 20% die from it. The exact conditions for causal links that convert risks to harms is an active area of research and current judicial investigations (60 Minutes, 2022; Ortutay, 2023; SMVLC, 2023).

8%-10% of algorithms are ‘bad’…

## 6. Conclusions

Analyzing 151 algorithmic audits from 33 studies, we found that around 8%-10% of algorithmic recommendations are 'bad', while a quarter ‘do good’. Algorithms drag users ‘out of the self-inflicted rabbit hole’ in an upward spiral more often (40% ‘good-to-good’ recommendation) than they lock them into them 11% ‘bad-to-bad recommendation’). The level of risk is surprisingly consistent across platforms from (YouTube, Google Search, Twitter, Facebook, TikTok, Amazon, etc.), and the type of harms (bias, mental health, misinformation, child harm, extremism, etc.), which justifies treating recommender algorithms as one class.

Within a larger context, we find that the detected risk and harms are similar to the levels of risk and harm found in generic food defects, which is a tightly monitored and regulated industry and providers are demanded to exhibit important levels of self-control and corporate responsibility. The analogy seems apt, because also here, most ‘apples’ are ‘good apples’, contributing positively to human health, and, as the saying goes, ‘an apple a day might keep the doctor away’. While it is important not to lose sight of this, the outstanding task seems to consists in picking out the ‘bad apples’, as even few of them ‘can spoil the bunch’.

By quantifying a ballpark measure of risks, our study contributes to the discussion of how to ensure the minimization of risks in a systematic way. In most other sectors of the economy dealing with products that carry some level of risk, the regular assessment of such is carried out by independent, public oversight agencies put in charge of regular audits. Such audits would be quite similar to the 151 academic algorithmic audits reviewed in this meta-analysis. Such agencies usualy also serve as a clearing house to facilitate and structure the ongoing dialogue about such risks. This includes discussions about tolerable and adequate levels of risks for different user segments, and the question if existing levels of risks are socially acceptable. This starts by quantifying risks and utilities, as done in this study.

**Author Contribution Statement:**
All authors (MH, AT, FJ, XZ, JYB, PB) contributed to the development of the data classification methodology, data collection, and data curation. MH conceptualized the study, contributed to the analysis, and wrote the article. AT contributed to the analysis. All authors reviewed and edited the article. MH, AT, FJ, PF, JYB contributed to the revise and resubmit process.

**Declaration of generative AI and AI-assisted technologies in the writing process**

During the preparation of this work the main author used ChatGPT-4 and Claude.AI to shorten some paragraphs of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

8%-10% of algorithms are ‘bad’…

## References

60 Minutes (Director). (2022, December 12). *Suing Social Media: Families say social media algorithms put their kids in danger | 60 Minutes*. https://www.youtube.com/watch?v=ItAseX1x_9o

Abul-Fottouh, D., Song, M. Y., & Gruzd, A. (2020). Examining algorithmic biases in YouTube’s recommendations of vaccine videos. *International Journal of Medical Informatics*, *140*, 104175. https://doi.org/10.1016/j.ijmedinf.2020.104175

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D’Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, *60*, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

Albadi, N., Kurdi, M., & Mishra, S. (2022). Deradicalizing YouTube: Characterization, Detection, and Personalization of Religiously Intolerant Arabic Videos. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 505:1-505:25. https://doi.org/10.1145/3555618

Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. (2021). Technologically scaffolded atypical cognition: The case of YouTube’s recommender system. *Synthese*, *199*(1), 835–858. https://doi.org/10.1007/s11229-020-02724-x

Alfonsi, S. (Director). (2022, December 11). *More than 1,200 families suing social media companies over kids’ mental health*. CBS News. https://www.cbsnews.com/news/social-media-lawsuit-meta-tiktok-facebook-instagram-60-minutes-2022-12-11/

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook’s Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The Welfare Effects of Social Media. *American Economic Review*, *110*(3), 629–676. https://doi.org/10.1257/aer.20190658

Attewell, R. G., Glase, K., & McFadden, M. (2001). Bicycle helmet efficacy: A meta-analysis. *Accident Analysis & Prevention*, *33*(3), 345–352. https://doi.org/10.1016/S0001-4575(00)00048-8

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Bandy, J., & Diakopoulos, N. (2021a). Curating quality? How Twitter’s timeline algorithm treats different types of news. *Social Media+ Society*, *7*(3), 20563051211041648.

Bandy, J., & Diakopoulos, N. (2021b). More Accounts, Fewer Links: How Algorithmic Curation Impacts Media Exposure in Twitter Timelines. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 78:1-78:28. https://doi.org/10.1145/3449152

Bennett, J., & Lanning, S. (2007). The netflix prize. *Proceedings of KDD Cup and Workshop*, *2007*, 35.

Berger, P. L., & Luckmann, T. (1967). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge* (First Thus). Anchor.

Bergman, M. P. (2023). Assaulting the Citadel of Section 230 Immunity: Products Liability, Social Media, and the Youth Mental Health Crisis. *Lewis & Clark Law Review*, *26*(4), 1159–1202.

Betsch, C., & Wicker, S. (2012). E-health use, vaccination knowledge and perception of own risk: Drivers of vaccination uptake in medical students. *Vaccine*, *30*(6), 1143–1148. https://doi.org/10.1016/j.vaccine.2011.12.021

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Braghieri, L., Levy, R., & Makarin, A. (2022). Social Media and Mental Health. *American Economic Review*, *112*(11), 3660–3693. https://doi.org/10.1257/aer.20211218

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

CCDH, (Center for Countering Digital Hate). (2022). *Deadly By Design: TikTok pushes harmful content promoting eating disorders and self-harm into users' feeds.* https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf

CDC. (2023a, January 10). *Road Traffic Injuries and Deaths—A Global Problem*. Centers for Disease Control and Prevention. https://www.cdc.gov/injury/features/global-road-safety/index.html

CDC, (Centers for Disease Control and Prevention). (2023b). *CDC report shows concerning increases in sadness and exposure to violence among teen girls and LGBQ+ youth* (CDC's Youth Risk Behavior Survey (YRBS)). https://www.cdc.gov/nchhstp/newsroom/fact-sheets/healthy-youth/sadness-and-violence-among-teen-girls-and-LGBQ-youth-factsheet.html

Chen, A. Y., Nyhan, B., Reifler, J., Robertson, R. E., & Wilson, C. (2022). Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos. In *arXiv e-prints*. https://doi.org/10.48550/arXiv.2204.10921

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*, 1–14.

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. https://doi.org/10.48550/arXiv.2307.09009

Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021). Neutral bots probe political bias on social media. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-25738-6

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Consumer Reports. (2023). *Takata Airbag Recall: Everything You Need to Know*. Consumer Reports. https://www.consumerreports.org/cars/car-recalls-defects/takata-airbag-recall-everything-you-need-to-know-a1060713669/

Culkin, J. M. (1967, March 18). A Schoolman’s Guide to Marshall McLuhan. *The Saturday Review*, 51–53, 66–79.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., & Sampath, D. (2010). The YouTube Video Recommendation System. *Proceedings of the Fourth ACM Conference on Recommender Systems*, 293–296. https://doi.org/10.1145/1864708.1864770

Dhir, A., Yossatorn, Y., Kaur, P., & Chen, S. (2018). Online social media fatigue and psychological wellbeing—A study of compulsive use, fear of missing out, fatigue, anxiety and depression. *International Journal of Information Management*, *40*, 141–152. https://doi.org/10.1016/j.ijinfomgt.2018.01.012

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Duffin, E. (2023). *Rate of identity theft reports, by state U.S. 2022*. Statista. https://www.statista.com/statistics/302370/rate-of-identity-theft-reports-in-the-us/

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Egger, M., & Smith, G. D. (1997). Meta-Analysis. Potentials and promise. *BMJ (Clinical Research Ed.)*, *315*(7119), 1371–1374. https://doi.org/10.1136/bmj.315.7119.1371

EricsCause. (2023). *Data and Victim Map*. Erik’s Cause. https://www.erikscause.org/maps_data

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Facebook. (2021). *Teen Mental Health Deep Dive (Oct 2019), with Facebook Annotation (Sept 2021)*. https://about.fb.com/wp-content/uploads/2021/09/Instagram-Teen-Annotated-Research-Deck-2.pdf

Faddoul, M., Chaslot, G., & Farid, H. (2020). *A Longitudinal Analysis of YouTube’s Promotion of Conspiracy Videos* (arXiv:2003.03318). arXiv. https://doi.org/10.48550/arXiv.2003.03318

FDA, (U.S. Food & Drug Administration). (2023a, January 24). FDA Announces Action Levels for Lead in Categories of Processed Baby Foods. *FDA Newsroom*. https://www.fda.gov/news-events/press-announcements/fda-announces-action-levels-lead-categories-processed-baby-foods

FDA, (U.S. Food & Drug Administration). (2023b, August 10). Closer to Zero: Reducing Childhood Exposure to Contaminants from Foods. *FDA Center for Food Safety and Applied Nutrition*. https://www.fda.gov/food/environmental-contaminants-food/closer-zero-reducing-childhood-exposure-contaminants-foods

Fletcher, R., & Nielsen, R. K. (2018). Automated Serendipity. *Digital Journalism*, *6*(8), 976–989. https://doi.org/10.1080/21670811.2018.1502045

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington Technical Report*, *2*(8).

Frost, R. L., & Rickwood, D. J. (2017). A systematic review of the mental health outcomes associated with Facebook use. *Computers in Human Behavior*, *76*, 576–600. https://doi.org/10.1016/j.chb.2017.08.001

Ganguli, D., Hernandez, D., Lovitt, L., DasSarma, N., Henighan, T., Jones, A., Joseph, N., Kernion, J., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Elhage, N., Showk, S. E., Fort, S., Hatfield-Dodds, Z., Johnston, S., … Clark, J. (2022). Predictability and Surprise in Large Generative Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764. https://doi.org/10.1145/3531146.3533229

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 2053951720943234. https://doi.org/10.1177/2053951720943234

Glasper, E. A. (2023). Is Social Media Fuelling Deaths Among Children? *Comprehensive Child and Adolescent Nursing*, *46*(1), 1–4. https://doi.org/10.1080/24694193.2023.2172291

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Haidt, J., Rausch, Z., & Twenge, J. (2023). Social media and mental health: A collaborative review. *Unpublished Manuscript, New York University. Retrieved from: Tinyurl. Com/SocialMediaMentalHealthReview*, 329.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (7th edition). Pearson.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125–2126. https://doi.org/10.1145/2939672.2945386

Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring Price Discrimination and Steering on E-commerce Web Sites. *Proceedings of the 14th ACM/USENIX Internet Measurement Conference (IMC’14)*. http://personalization.ccs.neu.edu/PriceDiscrimination/Research/

Hargreaves, E., Agosti, C., Menasché, D., Neglia, G., Reiffers-Masson, A., & Altman, E. (2018). Biases in the facebook news feed: A case study on the italian elections. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 806–812.

He, R., Lee, W. S., Ng, H. T., & Dahlmeier, D. (2017). An Unsupervised Neural Attention Model for Aspect Extraction. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 388–397. https://doi.org/10.18653/v1/P17-1036

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons.

Hilbert, M., Liu, B., Luu, J., & Fishbein, J. (2019). Behavioral Experiments With Social Algorithms: An Information Theoretic Approach to Input–Output Conversions. *Communication Methods and Measures*, *0*(0), 1–20. https://doi.org/10.1080/19312458.2019.1620712

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, *118*(32), e2101967118. https://doi.org/10.1073/pnas.2101967118

Hu, D., Jiang, S., E. Robertson, R., & Wilson, C. (2019). Auditing the partisanship of Google search snippets. *The World Wide Web Conference*, 693–704.

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 48:1-48:27. https://doi.org/10.1145/3392854

Hylton, K. N. (2012). The Law and Economics of Products Liability. *Notre Dame Law Review*, *88*(5), 2457–2514.

Jeyaraj, A., & Dwivedi, Y. K. (2020). Meta-analysis in information systems research: Review and recommendations. *International Journal of Information Management*, *55*, 102226. https://doi.org/10.1016/j.ijinfomgt.2020.102226

Ji-Xu, A., Htet, K. Z., & Leslie, K. S. (2023). Monkeypox Content on TikTok: Cross-sectional Analysis. *Journal of Medical Internet Research*, *25*, e44697.

Juneja, P., Bhuiyan, M. M., & Mitra, T. (2023). *Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube*. https://doi.org/10.1145/3544548.3580846

Juneja, P., & Mitra, T. (2021). Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–27. https://doi.org/10.1145/3411764.3445250

Kaplan, S., & Garrick, B. J. (1981). On The Quantitative Definition of Risk. *Risk Analysis*, *1*(1), 11–27. https://doi.org/10.1111/j.1539-6924.1981.tb01350.x

Kaushal, R., Saha, S., Bajaj, P., & Kumaraguru, P. (2016). KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube. *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, 157–164. https://doi.org/10.1109/PST.2016.7906950

Kranzberg, M. (1986). Technology and History: “Kranzberg’s Laws.” *Technology and Culture*, *27*(3), 544. https://doi.org/10.2307/3105385

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966–2981.

Lanier, J. (2018). *Ten Arguments for Deleting Your Social Media Accounts Right Now*. Henry Holt and Company.

Lazer, D., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lorenz, E. (1972). *Predictability: Does the flap of a butterfly’s wing in Brazil set off a tornado in Texas?*

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), Article 1. https://doi.org/10.1038/s41562-022-01460-1

Makhortykh, M., Urman, A., & Wijermars, M. (2022). A story of (non) compliance, bias, and conspiracies: How Google and Yandex represented Smart Voting during the 2021 parliamentary elections in Russia. *Harvard Kennedy School Misinformation Review*, *3*(2), 1–16.

McCrae, N., Gettings, S., & Purssell, E. (2017). Social Media and Depressive Symptoms in Childhood and Adolescence: A Systematic Review. *Adolescent Research Review*, *2*(4), 315–330. https://doi.org/10.1007/s40894-017-0053-4

Meier, A., & Reinecke, L. (2021). Computer-mediated communication, social media, and mental health: A conceptual and empirical meta-review. *Communication Research*, *48*(8), 1182–1209.

Memon, A. M., Sharma, S. G., Mohite, S. S., & Jain, S. (2018). The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. *Indian Journal of Psychiatry*, *60*(4), 384. https://doi.org/10.4103/psychiatry.IndianJPsychiatry_414_17

Merrer, E. L., & Trédan, G. (2022). *Surfing Personalization for Quantifying the Rabbit Hole Phenomenon on YouTube*. HAL open science. https://hal.science/hal-03620039

Mill, J. S. (2008). Utilitarianism. In *Seven Masterpieces of Philosophy*. Routledge.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Nisar, T. M., Prabhakar, G., Ilavarasan, P. V., & Baabdullah, A. M. (2019). Facebook usage and mental health: An empirical study of role of non-directional social comparisons in the UK. *International Journal of Information Management*, *48*, 53–62. https://doi.org/10.1016/j.ijinfomgt.2019.01.017

Nodder, C. (2013). *Evil by Design: Interaction Design to Lead Us into Temptation*. John Wiley & Sons.

NSC, (National Safety Council). (2023). *Injury Facts—National Safety Council*. Injury Facts. https://injuryfacts.nsc.org/

OpenAI. (2023). *GPT-4 System Card*. https://cdn.openai.com/papers/gpt-4-system-card.pdf

Orben, A. (2020). Teenagers, screens and social media: A narrative review of reviews and key studies. *Social Psychiatry and Psychiatric Epidemiology*, *55*(4), 407–414. https://doi.org/10.1007/s00127-019-01825-4

Orlowski, J. (Director). (2020). *The Social Dilemma* [Documentary]. Netflix. https://www.netflix.com/title/81254224

Ortutay, B. (2023, October 24). States sue Meta claiming its social platforms are addictive and harm children's mental health. *AP News*. https://apnews.com/article/instagram-facebook-children-teens-harms-lawsuit-attorney-general-1805492a38f7cee111cbb865cc786c28

Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., Stringhini, G., & Sirivianos, M. (2020). Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 522–533. https://doi.org/10.1609/icwsm.v14i1.7320

Papadamou, K., Zannettou, S., Blackburn, J., Cristofaro, E. D., Stringhini, G., & Sirivianos, M. (2022). "It Is Just a Flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 723–734. https://doi.org/10.1609/icwsm.v16i1.19329

Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2021). "How over is it?" Understanding the Incel Community on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 412:1-412:25. https://doi.org/10.1145/3479556

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin.

Parr, B. (2015). *Captivology: The Science of Capturing People's Attention*. Harper Collins.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Pasquali, M. (2022). *Fraudulent online order rate by business size 2021*. Statista. https://www.statista.com/statistics/1350621/share-fraudulent-online-orders-business-size/

Pearl, J. (2009). *Causality*. Cambridge University Press.

Petrosyan, A. (2022). *Experiences of online banking-related frauds by frequency 2019: Denmark, Sweden and Finland*. Statista. https://www.statista.com/statistics/498141/frequency-of-experiences-of-bank-card-and-online-banking-fraud-in-finland/; https://www.statista.com/statistics/498122/frequency-of-experiences-of-bank-card-and-online-banking-fraud-in-sweden/; https://www.statista.com/statistics/871230/frequency-of-experiences-of-bank-card-and-online-banking-fraud-in-denmark/

Petticrew, M., & Roberts, H. (2008). *Systematic Reviews in the Social Sciences: A Practical Guide*. John Wiley & Sons.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441.

Popper, K. R. (2002). *The Logic of Scientific Discovery*. Routledge.

Rahwan, I., & Cebrian, M. (2018, March 29). *Machine Behavior Needs to Be an Academic Discipline*. Nautilus. https://nautil.us/machine-behavior-needs-to-be-an-academic-discipline-237022/

8%-10% of algorithms are ‘bad’…

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477. https://doi.org/10.1038/s41586-019-1138-y

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Raychoudhury, P. (2021, September 26). What Our Research Really Says About Teen Well-Being and Instagram. *Meta*. https://about.fb.com/news/2021/09/research-teen-well-being-and-instagram/

Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Commun. ACM*, *40*(3), 56–58. https://doi.org/10.1145/245108.245121

Reuben, A., Caspi, A., Belsky, D. W., Broadbent, J., Harrington, H., Sugden, K., Houts, R. M., Ramrakha, S., Poulton, R., & Moffitt, T. E. (2017). Association of Childhood Blood Lead Levels With Cognitive Function and Socioeconomic Status at Age 38 Years and With IQ Change and Socioeconomic Mobility Between Childhood and Adulthood. *JAMA*, *317*(12), 1244–1251. https://doi.org/10.1001/jama.2017.1712

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (Eds.). (2011). *Recommender Systems Handbook*. Springer US. https://doi.org/10.1007/978-0-387-85820-3

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Ryan’s World. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ryan%27s_World&oldid=1145962866

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, *22*(2014), 4349–4357.

SCOTUS blog. (2023). Gonzalez v. Google LLC. *SCOTUSblog*. https://www.scotusblog.com/case-files/cases/gonzalez-v-google-llc/

Shin, D., Kee, K. F., & Shin, E. Y. (2022). Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? *International Journal of Information Management*, *65*, 102494. https://doi.org/10.1016/j.ijinfomgt.2022.102494

Shorten, A., & Shorten, B. (2013). What is meta-analysis? *Evidence-Based Nursing*, *16*(1), 3–4. https://doi.org/10.1136/eb-2012-101118

SMVLC, (Social Media Victims Law Center). (2023). *Social Media Addiction Lawsuits*. https://socialmediavictims.org/

Sohn, S. Y., Rees, P., Wildridge, B., Kalk, N. J., & Carter, B. (2019). Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: A systematic review, meta-analysis and GRADE of the evidence. *BMC Psychiatry*, *19*(1), 356. https://doi.org/10.1186/s12888-019-2350-x

Srba, I., Moro, R., Tomlein, M., Pecher, B., Simko, J., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., Gavornik, A., & Bielikova, M. (2022). Auditing YouTube’s Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Transactions on Recommender Systems*. https://doi.org/10.1145/3568392

Statista. (2023a). *Digital ad security violation rate 2022*. https://www.statista.com/statistics/1274304/digital-advertising-security-violation-rate-worldwide/

Statista. (2023b). *Global daily social media usage 2023*. Statista. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Stempel, J. (2020, March 27). GM reaches settlement over lost vehicle value from defective ignition switches. *Reuters*. https://www.reuters.com/article/us-gm-settlement-idUSKBN21E3LG

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., Beattie, L., Ekstrand, M., Leibowicz, C., Sehat, C. M., Johansen, S., Kerlin, L., Vickrey, D., Singh, S., Vrijenhoek, S., Zhang, A., Andrus, M., Helberger, N., Proutskova, P., … Vasan, N. (2022). *Building Human Values into Recommender Systems: An Interdisciplinary Synthesis* (arXiv:2207.10192). arXiv. https://doi.org/10.48550/arXiv.2207.10192

Sunstein, C. R. (2005). *Laws of fear: Beyond the precautionary principle* (pp. xii, 234). Cambridge University Press. https://doi.org/10.1017/CBO9780511790850

The New York Times (Director). (2020). *Rabbit Hole*. https://www.nytimes.com/column/rabbit-hole

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90–99. https://doi.org/10.1037/0033-2909.99.1.90

Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., & Bielikova, M. (2021). An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. *Proceedings of the 15th ACM Conference on Recommender Systems*, 1–11. https://doi.org/10.1145/3460231.3474241

Traynor, R. J. (1964). The ways and meanings of defective products and strict liability. *Tenn. L. Rev.*, *32*, 363.

Tutt, A. (2017). An FDA for Algorithms. *Administrative Law Review*, *69*(1), 83–124.

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2020). Considering All of the Data on Digital-Media Use and Depressive Symptoms: Response to Ophir, Lipshits-Braziler, and Rosenberg (2020). *Clinical Psychological Science*, *8*(2), 379–383. https://doi.org/10.1177/2167702619898179

Uman, L. S. (2011). Systematic Reviews and Meta-Analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *20*(1), 57–59.

Urman, A., Makhortykh, M., Ulloa, R., & Kulshrestha, J. (2022). Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and Informatics*, *72*, 101860. https://doi.org/10.1016/j.tele.2022.101860

Ustalov, D., Fedorova, N., & Pavlichenko, N. (2022). Improving Recommender Systems with Human-in-the-Loop. *Proceedings of the 16th ACM Conference on Recommender Systems*, 708–709. https://doi.org/10.1145/3523227.3547373

Uzogara, E. E. (2023). Democracy Intercepted. *Science*, *381*(6656), 386–387. https://doi.org/10.1126/science.adj7023

Valkenburg, P. M., & Peter, J. (2013). The Differential Susceptibility to Media Effects Model. *Journal of Communication*, *63*(2), 221–243. https://doi.org/10.1111/jcom.12024

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Williams, R., & Edge, D. (1996). The social shaping of technology. *Research Policy*, *25*(6), 865–899. https://doi.org/10.1016/0048-7333(96)00885-2

World of Coca Cola. (2023). *Vault of the secret formula*. https://www.worldofcoca-cola.com/explore-inside/explore-vault-secret-formula

Worldometer. (2023). *Worldometer—Real time world statistics*. Worldometer. http://www.worldometers.info/

Yesilada, M., & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, *11*(1). https://policyreview.info/articles/analysis/systematic-review-youtube-recommendations-and-problematic-content

Yoon, S., Kleinman, M., Mertz, J., & Brannick, M. (2019). Is social network site usage related to depression? A meta-analysis of Facebook–depression relations. *Journal of Affective Disorders*, *248*, 65–72. https://doi.org/10.1016/j.jad.2019.01.026

Yudkowsky, E. (2022). AGI Ruin: A List of Lethalities. *2022 MIRI Alignment Discussion*. https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities

Zade, H., Wack, M., Zhang, Y., Starbird, K., Calo, R., Young, J., & West, J. D. (2022). Auditing Google's Search Headlines as a Potential Gateway to Misleading Content: Evidence from the 2020 US Election. *Journal of Online Trust and Safety*, *1*(4), Article 4. https://doi.org/10.54501/jots.v1i4.72

# Supporting Information

## Contents

## S.I.1. Data repository

The data is publicly available and accessible at:

https://osf.io/gnm7h/?view_only=4408926c020940438a929ca7df59ecfb (anonymized link)

## YouTube studies

### (Kaushal et al., 2016) >>>>>

Kaushal, R., Saha, S., Bajaj, P., & Kumaraguru, P. (2016). KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube. *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, 157–164. https://doi.org/10.1109/PST.2016.7906950

**Type of risk:** kids. **Modality:** recommendations (API). **Main source:** Table V (p. 7).

**Collection method:** "For the purpose of our work, indecent videos are those which either have graphic nudity or abusive/inappropriate dialogues… [and] unsafe for viewing by kids… video-video communities based on these suggested (related) videos by using YouTube API… For each video, the algorithm finds top 10 related videos." (pages 1; 3; 6)

**Risk-Utility classification:** We classify 'unsafe' as 'bad', and 'safe' as 'other'.

8%-10% of algorithms are ‘bad’…

## (Papadamou et al., 2020) >>>>>

Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., Stringhini, G., & Sirivianos, M. (2020). Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 522–533. https://doi.org/10.1609/icwsm.v14i1.7320

**Type of risk:** misinformation. **Modality:** recommendations (API). **Main source:** Table 9 (p. 531).

**Collection method:** “we label these videos as one of four categories: 1) suitable; 2) disturbing; 3) restricted (equivalent to MPAA’s3 NC-17 and R categories); and 4) irrelevant videos… for the sake of our analysis …, we collapse our four labels into two general categories, by combining the suitable with the irrelevant videos into one “appropriate” category … and the disturbing with the restricted videos into a second “inappropriate” category… We call the first category “appropriate” despite including PG and PG-13 videos because those videos are not aimed at toddlers (irrelevant). On the other hand, videos rated as PG or PG-13 that target toddlers (aged 1 to 5) are disturbing and fall under the inappropriate category…. we iteratively collect the top 10 recommended videos… as returned by the YouTube Data API, for up to three hops within YouTube’s recommendation graph” (p. 523 * 530).

**Risk-Utility classification:** We classify ‘inappropriate’ as ‘bad’, ‘appropriate’ as ‘other’.

## (Merrer & Trédan, 2022) >>>>>

Le Merrer, E., & Trédan, G. (2022). Surfing Personalization for Quantifying the Rabbit Hole Phenomenon on YouTube. https://hal.science/hal-03620039/document

**Type of risk:** kids. **Modality:** autoplay. **Main source:** Fig. 1 (p. 5).

**Collection method:** “This study relies on bots following so called Autoplay recommendations…. When watching a video, the corresponding autoplay recommendation is the top recommendation displayed (right hand side of the YouTube video page), and is the next video to be played if no other action is taken by a user.” (p.3).

**Risk-Utility classification:** the authors start by explicitly watching a kids’ channel (which is assumed rated at G, general audience, unrestricted) and count the percentage of resulting ‘preschool’, ‘younger kids’, ‘older kids’ and ‘adult’ video recommendations after 10-20 autoplay loops. We classify ‘adult’ as ‘bad’, ‘preschool’ as ‘good’, which makes the assumption that it would be ‘bad’ for a recommender algorithm to guide someone from a kids channel to adult content.

## (Faddoul et al., 2020) >>>>>

Faddoul, M., Chaslot, G., & Farid, H. (2020). A longitudinal analysis of YouTube's promotion of conspiracy videos. arXiv https://doi.org/10.48550/arXiv.2003.03318

**Type of risk:** misinformation. **Modality:** autoplay. **Main source:** Figure 2 & 3 (p. 4-5).

**Collection method:** “We focus on the watch-next algorithm, which is the system that recommends a video to be shown next when auto-play is enabled.”

**Risk-Utility classification:** We classify ‘conspiracy’ as "bad", and the rest as ‘other’.

## (Alfano et al., 2021) >>>>>

Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. (2021). Technologically scaffolded atypical cognition: The case of YouTube’s recommender system. *Synthese*, *199*, 835-858.

8%-10% of algorithms are ‘bad’…

**Type of risk:** misinformation. **Modality:** search results. **Main source:** Table 3 (p. 850).

**Collection method:** “the robot starts with search query (q) results and obtains the first k videos that the YouTube search engine returns in response to the search query. Then, for every one of those videos, the robot collects the recommended videos and selects the top b recommendations recursively until it reaches the desired depth.” (p. 842)

**Risk-Utility classification:** we classify the ratio of conspiratorial clips for each category as 'bad' and the rest as other.

## (Papadamou et al., 2021) >>>>>

Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2021). "How over is it?" Understanding the Incel Community on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-25.

**Type of risk:** mental health. **Modality:** recommendations (API). **Main source:** Table 4 (p. 412:14).

**Collection method:** “To build the recommendation graphs used for our analysis, we use functionality provided by the YouTube Data API. For each video in the Incel-derived and Control sets, we collect the top 10 recommended videos associated with it.” (p. 412:13)

**Risk-Utility classification:** According to the authors, “Incels are one of the most extreme communities of the Manosphere…. Incel ideology is often associated with misogyny and anti-feminist viewpoints, and it has also been linked to multiple mass murders and violent offenses” (p. 1). We classify ‘incel-related videos’ as ‘bad’, and the rest as ‘other’.

## (Albadi et al., 2022) >>>>>

Albadi, N., Kurdi, M., & Mishra, S. (2022). Deradicalizing YouTube: Characterization, Detection, and Personalization of Religiously Intolerant Arabic Videos. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 505:1-505:25. https://doi.org/10.1145/3555618

**Type of risk:** political . **Modality:** recommendations (API) . **Main source:** Table 6 (p. 505:15).

**Collection method:** “We used YouTube’s API to collect data… collected through traversing five levels of recommendations to create a directed graph that resembled YouTube recommendations in which nodes represent videos and edges represent a recommendation activity” (p. 505:2, 505:15).

**Risk-Utility classification:** We classify ‘hateful’ as ‘bad’, and the rest as ‘other’.

## (Papadamou et al., 2022) >>>>>

Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2022, May). “It is just a flu”: Assessing the Effect of Watch History on YouTube’s Pseudoscientific Video Recommendations. *In Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 16, pp. 723-734).

**Type of risk:** misinformation . **Modality:** search results, YT home ro channel page, recommendations (API) . **Main source:** Table 4 (p. 732).

**Collection method:** “The YouTube Data API results are similar to those of the non-logged-in profile with no watch history (using a browser); this indicates that recommendations returned using the API are not subject to personalization… Homepage… We then visit each profile’s homepage to collect and classify the top 30 videos as ranked by YouTube… Search Results… We retrieve the top 20 videos for each search query… Video Recommendations… emulates the behavior of a user who watches videos based on recommendations, selecting the next video randomly from among the top 10 recommendations until he reaches five hops (p. 724; 728-729).

**Risk-Utility classification:** We classify 'scientific' as 'good', 'pseudo-scientific' as 'bad', 'sci/pseudo' as 'other' and 'browser' and 'data API' as 'neutral'.

## (Hosseinmardi et al., 2021) >>>>>

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences,* 118(32), e2101967118.

**Type of risk:** political . **Modality:** autoplay, search results, YT home or channel page . **Main source:** Tables 2, weighted by Table S2 (p. 6 and SI p. 3)

**Collection method:** "we used the YouTube API to retrieve the corresponding channel ID, as well as metadata such as the video's category, title, and duration… we explore how users encounter YouTube content by identifying 'referral' pages, which we define the page visited immediately prior to each YouTube video." (p. 2 and 6).

**Risk-Utility classification:** the authors define channels in six categories: far left (fL), left (L), center (C), AW (IDW/ASJW/MRA, grouped as 'anti-establishment'), right (R), and far right (fR). The authors see 'AW' not as explicitly extreme as 'far right', but rather as a kind of gateway to the far right. We classify 'far-left' and 'far-right' as 'bad', and 'left/right/center/ AW-anti-establishment' as 'other'.

## (Ribeiro et al., 2020) >>>>>

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr, W. (2020, January). Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 131-141). https://doi.org/10.1145/3351095.3372879

**Type of risk:** political . **Modality:** YT home or channel page . **Main source:** Table 3 (p. 138).

**Collection method:** "we devise a careful methodology to (a) collect a large pool of relevant channels; (b) collect data and the recommendations given by YouTube for these channel." (p. 133)

**Risk-Utility classification:** The authors see content linked to the Intellectual Dark Web (I.D.W.) as controversial without necessarily endorsing extreme views, while members of the Alt-right sponsor fringe ideas like that of a white ethnostate. Somewhere in the middle, individuals of the Alt-lite deny to embrace white supremacist ideology, although they frequently flirt with concepts associated with it (e.g., the "Great Replacement", globalist conspiracies). We classify mainstream 'media' as 'good', and all other three starting points as 'bad' (pulling extremer views 'out of the rabbit hole' into the mainstream). As for recommendations, we classify recommendations as 'bad' if they move farther to the right than the starting point, or to Alt-right (I.D.W. => Alt-lite or Alt-right is 'bad'; Alt-lite => Alt-right is 'bad'; Alt-right => Alt-right is 'bad'), and the rest as 'other'.

## (Srba et al., 2022) >>>>>

Srba, I., Moro, R., Tomlein, M., Pecher, B., Simko, J., Stefancova, E., ... & Bielikova, M. (2022). Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Transactions on Recommender Systems*.

**Type of risk:** misinformation . **Modality:** YT home or channel page, recommendations (side), search results . **Main source:** Fig. 3 (p. 6:29) and reports from p. 6:20.

8%-10% of algorithms are ‘bad’…

**Collection method:** “we let a series of agents (bots) pose as YouTube users. The agents performed predefined sequences of video watches and query searches. They also recorded items they saw: recommended videos, search results, and videos shown at the home page.” (p. 6:8)

**Risk-Utility classification:** We classify ‘misinformation-debunking’ as ‘good’, ‘misinformation-promoting’ as ‘bad’, and the rest as ‘other’.

## (Abul-Fottouh et al., 2020) >>>>>

Abul-Fottouh, D., Song, M. Y., & Gruzd, A. (2020). Examining algorithmic biases in YouTube’s recommendations of vaccine videos. *International Journal of Medical Informatics*, 140, 104175. https://doi.org/10.1016/j.ijmedinf.2020.104175

**Type of risk:** misinformation. **Modality:** recommendations (API). **Main source:** Table 1 (p. 4).

**Collection method:** “We used YouTube Data Tools… to retrieve YouTube videos related to vaccination and discover a network of recommended videos, in which videos are linked based on YouTube’s recommender algorithm. … For each video on this list, YouTube Data Tools retrieved up to 50 video recommendations provided by YouTube API.” (p. 2)

**Risk-Utility classification:** Starting with videos about vaccines, we follow governmental advice and classify ‘pro-vaccine’ as ‘good’, ‘anti-vaccine’ as ‘bad’, and the rest as ‘other’.

## (Hussein et al., 2020) >>>>>

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1-27. https://people.cs.vt.edu/tmitra/public/papers/cscw2020-YouTube_Audit.pdf

**Type of risk:** misinformation. **Modality:** autoplay, recommendations (side), search results. **Main source:** Fig. 6 (p. 48:19).

**Collection method:** “We collect the following components: (a) search results. These consist of top 20 videos in YouTube’s SERP (Search Engine Results Page) returned in response to a search query. (b) Up-Next corresponds to the next recommended video that will be played immediately after the current video finishes, (c) Top 5 relates to the top five recommended videos on the right of the video page. Figure 3 demonstrates the three components.” (p.48:10)

**Risk-Utility classification:** Starting out with conspiracy-prone topics, we classify ‘debunking’ as ‘good’, ‘promoting’ as ‘bad’, and the rest as ‘other’.

## (Tomlein et al., 2021) >>>>>>

Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., & Bielikova, M. (2021). An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. *Proceedings of the 15th ACM Conference on Recommender Systems*, 1–11. https://doi.org/10.1145/3460231.3474241

**Type of risk:** misinformation. **Modality:** recommendations (side) . **Main source:** Table 4 (p.12).

**Collection method:** “we annotated the following subset of collected videos: All recorded search results. Videos recommended for first 2 seed videos at the start of the run and last 2 seed videos of both phases (resulting in 6 sets of annotated videos per topic). This selection was a compromise between representativeness, correspondence to the reference study, and our capacities.We have not annotated the home page videos for the purpose of this study.” (p. 9)

**Risk-Utility classification:** we classify ‘conspiracy debunking’ as ‘good’, ‘conspiracy promoting’ as ‘bad’, and the rest as ‘other’. For starting conditions of anti-vaccine, we classify the start as neutral (S1), the promotion phase as bad (E1), and the debunking phase as good (E2).

## (A. Y. Chen et al., 2022) >>>>>

Chen, A. Y., Nyhan, B., Reifler, J., Robertson, R. E., & Wilson, C. (2022). Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos. arXiv preprint arXiv:2204.10921.

**Type of risk:** political. **Modality:** recommendations (side). **Main source:** Figure 8 (p. 20)

**Collection method:** “Participants voluntarily agreed to install a custom browser extension in Chrome or Firefox that monitored their web browsing behavior… our browser extension captures the exact videos that YouTube recommends” (p. 2-3). “...alternative channels tend to advocate “reactionary” positions and … combines Lewis’ Alternative Influence Network, the Intellectual Dark Web and Alt-lite channels from Ribeiro et al., and channels classified by Ledwich and Zaitsev as Men’s Rights Activists or Anti-Social Justice Warriors.”

**Risk-Utility classification:** We classify ‘mainstream’ as good, ‘extremist’ as ‘bad’, and recommendations toward more extremes as ‘bad’ (‘alternative’ to ‘extremist’ and ‘mainstream’ or ‘other’ to ‘alternative’ or ‘extremist’).

## (Juneja et al., 2023) >>>>>

Juneja, P., Bhuiyan, M. M., & Mitra, T. (2023). Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. arXiv preprint arXiv:2302.07836.

**Type of risk:** misinformation. **Modality:** autoplay. **Main source:** Fig. 11 (p. 14, “standard up-next trails”).

**Collection method:** “we designed a chrome browser extension named TubeCapture that enabled us to watch videos, conduct searches, and collect various YouTube components from users’ browsers” (p. 4-5).

**Risk-Utility classification:** We classify ‘supporting’ misinformation as ‘bad’, ‘opposing’ misinformation as ‘good’, and the rest as ‘neutral’.

## (Murthy, 2021) >>>>>

Murthy, D. (2021). Evaluating Platform Accountability: Terrorist Content on YouTube. *American Behavioral Scientist*, *65*(6), 800–824..

**Type of harm:** misinformation. **Modality:** recommendations (side). **Main source:** Fig. 2 (p. 813, “Frequency of videos in our sample by YouTube category.”).

**Collection method:** “11 videos were used as seeds to build a network that represents videos: (a) recommended, (b) recommended by the recommended videos, and (c) videos recommended by recommended videos in (b). Data was stored as a directional network and these were visualized using the SNA package Gephi to determine what ISIS videos were being recommended.” (p. 4-5).

8%-10% of algorithms are 'bad'…

**Risk-Utility classification:** We classify ISIS videos as 'bad', 'and the rest as 'others'.

## Search Engines studies

### (Urman et al., 2022) >>>>>

Urman, A., Makhortykh, M., Ulloa, R., & Kulshrestha, J. (2022). Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and informatics*, 72, 101860.

**Type of risk:** misinformation. **Modality:** search results. **Main source:** Figure 1 (p. 6)

**Collection method:** "we extracted all unique URLs that appeared on the first page of search results for each engine… Debunks conspiracy: the content under the URL lists argument(s) why a conspiracy theory is not true.... Promotes conspiracy: the content under the URL lists argument(s) in favor of a conspiracy theory... Mentions conspiracy: a conspiracy theory is mentioned, but there is no clear stance associated" (p. 4-5).

**Risk-Utility classification:** We classify 'debunked' as 'good', 'promotion' as 'bad', and 'no mention' or 'simple mention' as 'neutral'.

### (Makhortykh et al., 2022) >>>>>

Makhortykh, M., Urman, A., & Wijermars, M. (2022). A story of (non) compliance, bias, and conspiracies: How Google and Yandex represented Smart Voting during the 2021 parliamentary elections in Russia. *Harvard Kennedy School Misinformation Review*, 3(2), 1-16.

**Type of risk:** misinformation. **Modality:** search results. **Main source:** Figures 4-5 (p. 8-9)

**Collection method:** "Shares of content with different stance towards Smart Voting returned" (p. 9).

**Risk-Utility classification:** We classify 'debunked' as 'good', 'promotion' as 'bad', and 'no mention' or 'simple mention' as 'neutral'.

### (Zade et al., 2022) >>>>>

Zade, H., Wack, M., Zhang, Y., Starbird, K., Calo, R., Young, J., & West, J. D. (2022). Auditing Google's Search Headlines as a Potential Gateway to Misleading Content: Evidence from the 2020 US Election. *Journal of Online Trust and Safety*, 1(4).

**Type of risk:** misinformation. **Modality:** search results. **Main source:** Figures 3 and 6 (p. 14 and 17).

**Collection method:** "we used the SerpApi platform (SerpApi 2020) to search for keywords of our choice … and fetched the corresponding Google Search results as it would be seen at the 20 unique locations…" (p. 8).

**Risk-Utility classification:** We classify 'promotes trust' as 'good', 'promoted distrust' as 'bad', and rest as 'neutral'.

### (L. Chen et al., 2018) >>>>>

8%-10% of algorithms are ‘bad’…

Chen, L., Ma, R., Hannak, A, & Wilson, C. (2018) Investigating the Impact of Gender on Rank in Resume Search Engines | *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Retrieved January 25, 2023, from https://dl.acm.org/doi/10.1145/3173574.3174225

**Type of risk:** bias. **Modality:** search results. **Main source:** Table 6 (p. 8)

**Collection method:** “To collect data for this study, we use an automated web browser to search for candidates on Indeed, Monster, and CareerBuilder. Intuitively, our crawler is designed to emulate how a recruiter would search for candidates on these hiring websites. We ran queries for 35 job titles in 20 U. S. cities (described below) on all three sites, and recorded the resulting lists of candidates. We also queried for a subset of 490, 700, and 700 job title/city pairs with one, two, and three search filters, on Monster, Indeed, and CareerBuilder, respectively.” (p. 4).

**Risk-Utility classification:** We classify job title/city pairs that exhibit significant group unfairness as ‘bad’ and present the average over all reported job titles.

## (Hu et al., 2019) >>>>>

Hu, D., Jiang, S., E. Robertson, R., & Wilson, C. (2019, May). Auditing the partisanship of Google search snippets. *The World Wide Web Conference* (pp. 693-704). https://cbw.sh/static/pdf/hu-www19.pdf

**Type of risk:** political. **Modality:** search results. **Main source:** Figure 7 (p. 8)

**Collection method:** “We queried Google Search… We also queried Google Search for all of the autocomplete suggestions” (p. 4)

**Risk-Utility classification:** We classify content into three groups: unchanged partisanship, which we classify as ‘neutral’; pro- & counter partisanship (sum of ‘decrease’ and ‘flip’, the two counter-views, which we balance with the same amount of pro-views), which we classify as ‘other’; and then we calculate the partisan bias, which we define as the left-over pro-view that exceeds the offered counter-views. We classify this last group as ‘bad’ and the rest as ‘other’.

## (Betsch & Wicker, 2012) >>>>>>

Betsch, C., & Wicker, S. (2012). E-health use, vaccination knowledge and perception of own risk: drivers of vaccination uptake in medical students. *Vaccine*, 30(6), 1143-1148.

**Type of risk:** misinformation. **Modality:** search results. **Main source:** Table 5 (p. 1147)

**Collection method:** “we ‘googled’ the search terms … we categorized the first ten hits according to their source” (p. 1146)

**Risk-Utility classification:** We classify ‘correct’ info as ‘good’ and ‘false answers’ as ‘bad’, and the rest as other.

### TikTok

## (Ji-Xu et al., 2023) >>>>>

Ji-Xu, A., Htet, K. Z., & Leslie, K. S. (2023). Monkeypox Content on TikTok: Cross-sectional Analysis. *Journal of Medical Internet Research*, *25*, e44697.

**Type of risk:** misinformation. **Modality:** autoplay. **Main source:** Table 1 (p. 2)

8%-10% of algorithms are 'bad'…

**Collection method:** "TikTok was searched for videos using the term 'monkeypox'.... The top [100] videos returned by the TikTok search algorithm were analyzed." (p. 1)

**Risk-Utility classification:** We classify videos with misleading statements on monkeypox as 'bad', and the rest as other.

## (CCDH, 2022) >>>>>

CCDH (Center for Countering Digital Hate). (2022). Deadly By Design: TikTok pushes harmful content promoting eating disorders and self-harm into users' feeds. https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf

**Type of risk:** mental health. **Modality:** autoplay. **Main source:** Tables p. 14; 19; 25; 35.

**Collection method:** "recordings of the content served by TikTok's recommendation algorithm to new accounts in the first 30 minutes that they spend browsing the platform's 'For You' feed… [for] four accounts… researchers identified a recommended video matching one of the …[identified] categories, they viewed the video for 10 seconds and liked it" (p. 9-10). We also contacted Callum Hood, CCDH's Head of Research, who clarified that across "the four hours of recordings we made (comprising eight 30 minute recordings), we viewed 3,848 videos in total. So on average, we viewed videos for just 3.7 seconds [on average]. We stayed on relevant videos (videos fitting the categories of body image, mental health, self-harm/suicide or eating disorders) for at least 10 seconds and like the videos. So some videos will have been skipped much more quickly than 3.7 seconds and others will have been 10 seconds or a little more" (email communication 5/4/2023).

**Risk-Utility classification:** We translated the average watchtime of 3.7 seconds/video into frequencies to obtain the percentages of our framework. For the source, we classify the 'self-harm and suicide' and 'eating disorder' profiles as 'bad', and 'mental health' and 'body image' as other.

### Facebook

## (Bakshy et al., 2015) >>>>>

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130-1132. https://www.science.org/doi/10.1126/science.aaa1160

**Type of risk:** political. **Modality:** recommendations (API). **Main source:** Tables S6 and S7 (p. 22-23 Supplementary Material).

**Collection method:** Using Facebook internal data "We constructed a deidentified data set that includes 10.1 million active U.S. users who self-report their ideological affiliation" (p. 1130).

**Risk-Utility classification:** We classify the percentage of content that the algorithm filters out as 'bad' and the remaining content as neutral.

## (Ali et al., 2019) >>>>>

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction,* 3(CSCW), 199:1-199:30. https://doi.org/10.1145/3359301

**Type of risk:** bias. **Modality:** recommendations (newsfeed). **Main source:** Fig. 8 (p. 199:21).

8%-10% of algorithms are 'bad'…

**Collection method:** "we advertise eleven different generic job types… we customize the text, headline, and image as a real employment ad would…. We run these ads for 24 hours… showing a breakdown of ad delivery by gender and race in the ultimate delivery audience" (p. 199:20-21).

**Risk-Utility classification:** We classify the bias calculated as the average deviation from the expected value of the fraction of men/white users in the audience, as 'bad'.

## (Hargreaves et al., 2018) >>>>>

Hargreaves, E., Agosti, C., Menasché, D., Neglia, G., Reiffers-Masson, A., & Altman, E. (2018, August). Biases in the facebook news feed: a case study on the Italian elections. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 806-812). IEEE.

**Type of risk:** politial. **Modality:** recommendations (newsfeed). **Main source:** Fig. 1 (p. 4).

**Collection method:** "The collected snapshots and impressions, together with the set of all posts published by the thirty representative pages, obtained through Facebook API, constitute our dataset." (p. 3).

**Risk-Utility classification:** We classify as 'bad' when 'undecided', 'right' or 'far-right' get presented with 'far-right' content, and the rest as 'other'.

### Twitter

## (Bandy & Diakopoulos, 2021b) >>>>>

Bandy, J., & Diakopoulos, N. (2021). More accounts, fewer links: How algorithmic curation impacts media exposure in Twitter timelines. *Proc. of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-28.

**Type of risk:** political. **Modality:** recommendations (newsfeed). **Main source:** Fig. 5-6 (p. 78:18-19).

**Collection method:** "we followed a 'sock puppet' audit method…, whereby we identified a panel of users, simulated their interactions with the algorithmic Twitter timeline, and made comparisons to a baseline chronological timeline" (p. 78:6).

**Risk-Utility classification:** We classify as 'bad' as the increase of 'algorithmic' minus 'chronological' content and as the increase of 'injected' minus 'chronological' content of the same leaning (increase in left content for left-leaning and increase of right content for right-leaning).

## (Bandy & Diakopoulos, 2021a) >>>>>

Bandy, J., & Diakopoulos, N. (2021). Curating quality? How Twitter's timeline algorithm treats different types of news. *Social Media+ Society*, *7*(3), 20563051211041648.

**Type of risk:** political. **Modality:** recommendations (newsfeed). **Main source:** Table 3 (p. 9).

**Collection method:** "We collected timeline data twice per day for four full weeks" (p. 5).

**Risk-Utility classification:** We classify as 'bad' the percentage increase of conspiracy junk news in the algorithmic feed over the chronological feed.

## (W. Chen et al., 2021) >>>>>

Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021). Neutral bots probe political bias on social media. *Nature Communications*, 12(1), Article 1. https://doi.org/10.1038/s41467-021-25738-6.

8%-10% of algorithms are ‘bad’…

**Type of risk:** political. **Modality:** recommendation (newfeed). **Main source:** Figure 4 (p. 24).

**Collection method:** “We consider a list of low-credibility sources that are known to publish false and misleading news reports, conspiracy theories, junk science, and other types of misinformation” (p. 6).

**Risk-Utility classification:** We classify as ‘bad’ exposure to low-credibility content.

### Amazon

## (Juneja & Mitra, 2021) >>>>>

Juneja, P., & Mitra, T. (2021). Auditing e-commerce platforms for algorithmically curated vaccine misinformation. *Proceedings of the 2021 chi conference on human factors in computing systems* (1-27).

**Type of risk:** misinformation. **Modality:** recommendations (side), search results. **Main source:** Figure 6 (p. 13_.

**Collection method:** For RQ1 of Figure 6, “we collect search results without logging in to Amazon to eliminate the influence of personalization. …our Unpersonalized audits ran for 15 consecutive days, sorting the search results across 5 different Amazon filters each day: ‘featured’, ‘price low to high’, ‘price high to low’, ‘average customer review’ and ‘newest arrivals’” (p. 2).

**Risk-Utility classification:** We classify ‘debunking’ as ‘good’, ‘promoting’ as ‘bad’, and the rest as ‘other’.

### MIX (Facebook, Search Engine, Instagram, Twitter)

## (Lambrecht & Tucker, 2019) >>>>>

Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, 65(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

**Type of risk:** bias. **Modality:** search results. **Main source:** Table 2 (p. 2971); Table 8, 9 and 10 (p. 2977).

**Collection method:** “We ran advertising campaigns that directed users who clicked on the ad to [a] website… ‘Impressions’ refers to the number of times a particular ad was shown” (p. 2969-2970).

**Risk-Utility classification:** We classify as ‘bad’ the bias calculated by the difference of gender skewed impressions from the expected value (which is 50%).

8%-10% of algorithms are ‘bad’…

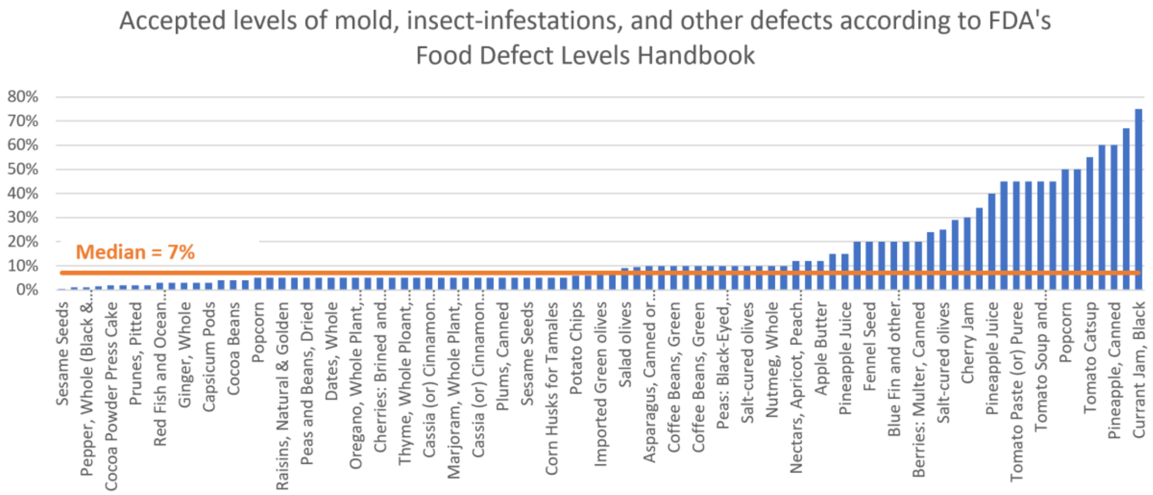
## S.I.2. Comparison data

In this section we present the reasoning behind the statistics we gathered in order to provide some context to the recommender algorithm risk levels we gathered. It is important to emphasize that these rough estimates are very much ‘back-of-the-envelope’ estimates and that we are not experts in these fields. These contextual estimates are not the main focus of our efforts and we are very happy for any constructive feedback on the numbers and our reasoning that we transparently present here.

### Food safety: general food defects

**Risk:** The U.S. Food and Drug Administration (FDA) maintains a Food Defect Levels Handbook, which outlines the “levels of natural or unavoidable defects in foods that present no health hazards for humans”.[8] The handbook lists over 111 products with a diverse list of defects (e.g. number of rodent hair per grams; and number of larvae per 500 grams, etc.). Aiming to have some ballpark number, we extracted the acceptable cases of mold, insect-infestations and related effects, which applies to 89 of the listed products. We find that the medium number of **accepted levels of mold, insect-infestations and related effects is at 7%**.

Figure S.I.1



**Harm:** The U.S. Centers for Disease Control and Prevention (CDC) estimates the “annual number of episodes of domestically acquired, foodborne illness, hospitalizations, and deaths caused by 31 pathogens and unspecified agents transmitted through food” in the U.S. to affect **15% of the U.S. population in terms of reported cases, 0.04% of the population being hospitalized and 0.0009% dying**.[9]

[8] https://www.fda.gov/food/ingredients-additives-gras-packaging-guidance-documents-regulatory-information/food-defect-levels-handbook

[9] Scallan, E., Griffin, P. M., Angulo, F. J., Tauxe, R., & Hoekstra, R. M. (n.d.). Foodborne Illness Acquired in the United States—Unspecified Agents—Volume 17, Number 1—January 2011—Emerging Infectious Diseases journal—CDC. https://doi.org/10.3201/eid1701.p21101

8%-10% of algorithms are 'bad'…

## Food safety: salmonella

**Risk:** The Food Safety and Inspection Service (FSIS) of the U.S. Department of Agriculture (USDA) has set performance standards that establish the maximum number of Salmonella-positive samples acceptable per sample set. These are mainly set at **7.5% of salmonella-positive samples** (i.e. for broiler chicken carcasses and ground beef), but can be as low as 1.7% (for young turkey carcasses). The agency classifies establishments into Category 1 if the salmonella-positive rate is at 3.75% or less, and Category 2 if it is between 3.75% and 7.5%.[10]

Scholarly studies detect a higher general rate outside the U.S. and rate that notably decreased in the U.S. in recent years, in general varying from **1% to 29% of salmonella positive samples**:

29% for retail chicken South Wales (2000);[11]
26% for the chicken carcasses Costa Rica (2013);[12]
17% for very small chicken establishments in the U.S. (2014);[13]
11% for chicken broiler production the U.S. (1990);[14]
7.2 % for the total U.S. in 2009;[13]
3.8% for the total U.S. in 2014; [13]
1.1% for large chicken establishments in the U.S. (2014); [13] and
1.% for the average in Morocco (2002-2005).[15]

**Harm:** The U.S. Centers for Disease Control and Prevention (CDC) estimates Salmonella bacteria to cause about 1.35M infections, 26,500 hospitalizations, and 420 deaths in the U.S. every year. This is equivalent to **0.4% of infections per inhabitant, 0.008% hospitalizations, and 0.0001% deaths** in the U.S. per year.

## Food safety: listeria monocytogenes

**Risk:** Since the late 1990s, "public health and regulatory agencies in the U.S. have established a zero tolerance for Listeria monocytogenes in cooked, ready-to-eat food".[16] Still, scholarly studies detect the bacterium that can cause a serious foodborne illness called listeriosis frequently, even if its prevalence

[10] https://www.fsis.usda.gov/sites/default/files/media_file/2021-02/10250.1.pdf

[11] Harrison, W. A., Griffith, C. J., Tennant, D., & Peters, A. C. (2001). Incidence of Campylobacter and Salmonella isolated from retail chicken and associated packaging in South Wales. Letters in Applied Microbiology, 33(6), 450–454. https://doi.org/10.1046/j.1472-765X.2001.01031.x

[12] Rivera-Pérez, W., Barquero-Calvo, E., & Zamora-Sanabria, R. (2014). Salmonella Contamination Risk Points in Broiler Carcasses during Slaughter Line Processing. Journal of Food Protection, 77(12), 2031–2034. https://doi.org/10.4315/0362-028X.JFP-14-052

[13] Thames, H. T., & Theradiyil Sukumaran, A. (2020). A Review of Salmonella and Campylobacter in Broiler Meat: Emerging Challenges and Food Safety Measures. Foods, 9(6), 776. https://doi.org/10.3390/foods9060776
and
www.fsis.usda.gov/sites/default/files/media_file/2021-02/Progress-Report-Salmonella-Campylobacter-CY2014.pdf

[14] Jones, F. T., Axtell, R. C., Rives, D. V., Scheideler, S. E., Tarver, F. R., Walker, R. L., & Wineland, M. J. (1991). A Survey of Salmonella Contamination in Modern Broiler Production. Journal of Food Protection, 54(7), 502–513. https://doi.org/10.4315/0362-028X-54.7.502

[15] Bouchrif, B., Paglietti, B., Murgia, M., Piana, A., Cohen, N., Ennaji, M. M., Rubino, S., & Timinouni, M. (2009). Prevalence and antibiotic-resistance of Salmonella isolated from food in Morocco. The Journal of Infection in Developing Countries, 3(01), Article 01. https://doi.org/10.3855/jidc.103

[16] Shank, F. R., Elliot, E. L., Wachsmuth, I. K., & Losikoff, M. E. (1996). US position on Listeria monocytogenes in foods. Food Control, 7(4), 229–234. https://doi.org/10.1016/S0956-7135(96)00041-2

has notably been reduced. In the late 1980s, it was detected in some 25% of samples.[17] Studies in Eastern countries found it in **12% of samples** between 2005 and 2017.[18] A study in Portugal found it in **7% of samples** in 2004.[19]

**Harm:** The U.S. Centers for Disease Control and Prevention (CDC) estimates 1,600 people get listeriosis each year (94% resulting in hospitalizations), and about 260 die.[20] This is equivalent to **0.001% reported cases per inhabitant, 0.0005% hospitalizations, and 0.0001% deaths**.

## Consumer goods

**Risk:** According to the U.S. Dept of Commerce, the "U.S. consumer market (defined as packaged goods)" was at $635 billion in 2019, and the same underlying source estimates $806 billion in 2023.[21] There are no statistics about the average cost per consumer product, but we asked a large language model (LLM: ChatGPT-4) to give us a generic estimate and then fed it with a list of common consumer products that we obtained from the U.S. Consumer Product Safety Commission (CPSC) and asked the LLM to calculate an estimated cost average, to get a better idea. After fine-tuning the prompts, the LLM suggests an average price of $25. This suggests that some 32 billion consumer products were sold in 2023. CPSC reports to have recalled 46.6 million products in 2022.[22] Hence, we estimate that [46.6M / ($806B/$25)] = **0.1% of products were recalled**.

**Harm:** The U.S. National Safety Council (NSC), based on CPSC National Electronic Injury Surveillance System (NEISS) data[23], estimates for 2022 some 14.3M consumer product-related injuries treated in hospital emergency departments.[24] The most dangerous products include home structures (i.e. stairs/ramps and tools) and recreation equipment (bicycles and exercise equipment). This suggests that there were 4.3 injuries per 100 U.S. inhabitants (U.S. population of 332M), which could suggests that **between 2-4 injuries per 100 people**.[25] Note that the number of recalls from above refers to a given year, while such injuries can arise from accumulated consumer products. The low single digit number

[17] Johnson, J. L., Doyle, M. P., & Cassens, R. G. (1990). Listeria monocytogenes and Other Listeria spp. In Meat and Meat Products A Review. Journal of Food Protection, 53(1), 81–91. https://doi.org/10.4315/0362-028X-53.1.81

[18] Jamshidi, A., & Zeinali, T. (2019). Significance and Characteristics of Listeria monocytogenes in Poultry Products. International Journal of Food Science, 2019, e7835253. https://doi.org/10.1155/2019/7835253

[19] Mena, C., Almeida, G., Carneiro, L., Teixeira, P., Hogg, T., & Gibbs, P. A. (2004). Incidence of Listeria monocytogenes in different food products commercialized in Portugal. Food Microbiology, 21(2), 213–216. https://doi.org/10.1016/S0740-0020(03)00057-1

[20] https://www.cdc.gov/listeria/index.html

[21] www.trade.gov/selectusa-consumer-goods-industry & www.statista.com/outlook/io/manufacturing/consumer-goods/united-states

[22] https://www.cpsc.gov/Data/Monthly-Progress-Reports This includes 37M low cost items, like Clorox cleaner liquid for some $5-$10, but also 2.3M automatic baby swings for $200, and 1.9M cooler cases for $300, and 1.3M miter saws for $350 a piece. This would suggest a higher average price per product than $25, but we expect that the probability of recalls increases with the size and complexity of a product, as compared to the many small products sold.

[23] https://injuryfacts.nsc.org/

[24] https://injuryfacts.nsc.org/home-and-community/safety-topics/consumer-product-injuries/data-details/

[25] The data system allows for reporting of up to two products for each person's injury, so a person's injury may be counted in two product groups.

agrees with the estimate that around 1% of the population reports having been victim to consumer fraud in a given year in the U.S.[26].

## Sports: soccer

**Risk:** The International Centre for Sports Studies (CIES)’s Football Observatory estimates an average of 26 fouls per game in European professional soccer leagues.[27] A study on the German Bundesliga measured that there are on average 836 “individual ball possession” intervals in a game.[28] This suggests an average rate of **3.1% of fouls per ball possession** (or, every 33rd possession is a foul, which sounds reasonable to a soccer player).

**Harm:** The U.S. National Safety Council (NSC), based on CPSC National Electronic Injury Surveillance System (NEISS) data[23], estimates for 2022 some 179,284 soccer injuries treated in emergency departments (of which 4,693 were hospitalized or dead-on-arrival). The U.S. Sports and Fitness Industry Association (SFIA) estimates some 11.9M active soccer (outdoor) players in the U.S..[29] This suggests some **1.5% injuries per active player** (this average increases to 3.5% for 6-24 year old).

## Health: smoking

**Risk:** The risks associated with smoking are produced by more ingredients than the well-known nicotine and tar. According to the American Lung Association, “when burned, cigarettes create more than 7,000 chemicals. At least 69 of these chemicals are known to cause cancer, and many are toxic. Many of these chemicals also are found in consumer products, but these products have warning labels—such as rat poison packaging. While the public is warned about the danger of the poisons in these products, there is no such warning for the toxins in tobacco smoke.”[30] We conclude that there is nothing like a harmless cigarette and that **100% of cigarettes pose a risk to human health**.

According to the World Health Organization (WHO), in 2020, **22.3% of the world’s population used tobacco** (with a skewed underlying distribution of 37% smokers being male and 8% female, and 80% living in low-and middle-income countries).[31] We consider this ‘risk’ of social attraction to smoking as the risk level. In the U.S., due to public campaigns, regulations and restrictions, smoking has declined from 20.9% in 2005 to 11.5% in 2021.[32] Globally, some 10 billion cigarettes are smoked per day. This results in some 6 cigarettes per smoker per day.

**Harm:** While smoking affects all kinds of health aspects (from mental health to all kinds of organs and tissue), we focus on cardiovascular- and lung-disease, which are the most evident and severe effects. We remove the baseline of the non-smoking population affected by the same conditions and calculate the remaining percentage of smokers affected who otherwise would not have been expected to be affected if they were not smoking smokers (traditionally one would calculate the odds-ratio here, but

[26] Duffin, E. (2023). Consumer fraud report rate, by state U.S. 2022. Statista. https://www.statista.com/statistics/302313/consumer-fraud-report-rate-in-the-us/

[27] https://football-observatory.com/IMG/sites/b5wp/2018/255/en/

[28] Link, D., & Hoernig, M. (2017). Individual ball possession in soccer. PLoS ONE, 12(7), e0179953. https://doi.org/10.1371/journal.pone.0179953

[29] https://sfia.medium.com/soccer-participation-in-the-united-states-92f8393f6469

[30] https://www.lung.org/quit-smoking/smoking-facts/whats-in-a-cigarette

[31] https://www.who.int/news-room/fact-sheets/detail/tobacco

[32] https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

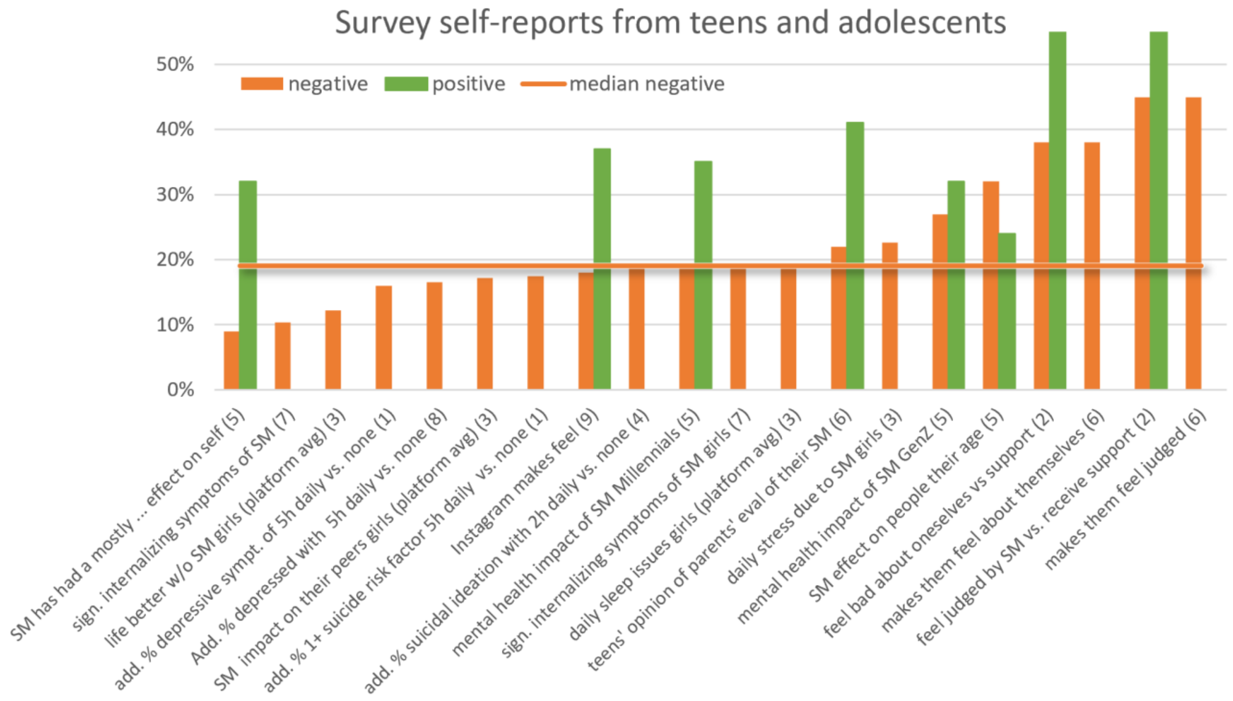
we want to keep it transparently comparable with simple percentages). The resulting percentage of unexpectedly impacted smokers varies between 19% and 2.5% for different diseases, with some **12% being the rough median for the most prominent complications** (cardiovascular events and lung cancer):

18.9% die additionally from all kinds of other non-cardiovascular disease;
12.4% have additional cardiovascular events;
11.7% have additional lung cancer;
8.7% have additional cardiovascular induced deaths;
7.8% have additional fatal or nonfatal myocardial infarction;
2.6% have additional heart failure; and
2.4% have additional fata or nonfatal strokes.[33]

## Self-reported social media harms by adolescents

We undertook a simple inventory of studies that report survey results of self-reported social media harms among adolescents. We found 20 articles through the very useful 330-page long collaborative review from Haidt, Rausch, and Twenge (Haidt et al., 2023). This rudimentary inventory was not the focus of our study and merely serves to provide a ballpark measure for contextual comparison.

Figure S.I.2



[33] Khan, S. S., Ning, H., Sinha, A., Wilkins, J., Allen, N. B., Vu, T. H. T., Berry, J. D., Lloyd-Jones, D. M., & Sweis, R. (2021). Cigarette Smoking and Competing Risks for Fatal and Nonfatal Cardiovascular Disease Subtypes Across the Life Course. Journal of the American Heart Association, 10(23), e021751. https://doi.org/10.1161/JAHA.121.021751
Bruder, C., Bulliard, J.-L., Germann, S., Konzelmann, I., Bochud, M., Leyvraz, M., & Chiolero, A. (2018). Estimating lifetime and 10-year risk of lung cancer. Preventive Medicine Reports, 11, 125–130. https://doi.org/10.1016/j.pmedr.2018.06.010

8%-10% of algorithms are ‘bad’…

(1) journals.sagepub.com/doi/abs/10.1177/2167702619898179?journalCode=cpxa
(2) www.apa.org/news/press/releases/stress/2018/stress-gen-z.pdf
(3) www.commonsensemedia.org/research/teens-and-mental-health-how-girls-really-feel-about-social-media
(4) www.liebertpub.com/doi/full/10.1089/cyber.2015.0055
(5) www.mckinsey.com/mhi/our-insights/gen-z-mental-health-the-impact-of-tech-and-social-media
(6) www.pewresearch.org/internet/2022/11/16/connection-creativity-and-drama-teen-life-on-social-media-in-2022/
(7) www.researchsquare.com/article/rs-2790469/v1
(8) www.thelancet.com/journals/eclinm/article/PIIS2589-5370(18)30060-9/fulltext
(9) www.wsj.com/articles/the-facebook-files-11631713039?mod=bigtop-breadcrumb

The median of the self-reported negative effect is at 19%. In line with our opening example from the Facebook papers, the self-reported positive effect is notably higher, with a median at 36%. In studies, positive effects are less frequently reported than negative effects. We find the same in our review of recommender algorithms.