Helping Users Tackle Algorithmic Threats on Social Media: A Multimedia Research Agenda

Christian von der Weth Ashraf Abdul Shaojing Fan Mohan Kankanhalli
School of Computing, National University of Singapore
{chris|ashraf|fansj|mohan}@comp.nus.edu.sg

ABSTRACT

Participation on social media platforms has many benefits but also poses substantial threats. Users often face an unintended loss of privacy, are bombarded with mis-/disinformation, or are trapped in filter bubbles due to over-personalized content. These threats are further exacerbated by the rise of hidden AI-driven algorithms working behind the scenes to shape users' thoughts, attitudes, and behaviour. We investigate how multimedia researchers can help tackle these problems to level the playing field for social media users. We perform a comprehensive survey of algorithmic threats on social media and use it as a lens to set a challenging but important research agenda for effective and real-time user nudging. We further implement a conceptual prototype and evaluate it with experts to supplement our research agenda. This paper calls for solutions that combat the algorithmic threats on social media by utilizing machine learning and multimedia content analysis techniques but in a transparent manner and for the benefit of the users.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

social media, privacy, fake news, echo chambers, user nudging, machine learning, explainable AI

ACM Reference Format:
Christian von der Weth Ashraf Abdul Shaojing Fan Mohan Kankanhalli. 2020. Helping Users Tackle Algorithmic Threats on Social Media: A Multimedia Research Agenda. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3414692

1 INTRODUCTION

As platforms for socializing but also as source of information, social media has become one of the most popular services on the Web. However, the use of social media also poses substantial threats to its users. The most prominent threats include a loss of privacy, mis-/disinformation (e.g., "fake news", rumors, hoaxes), and "over-personalized" content resulting in so-called filter bubbles and echo chambers. While these threats are not new, the risk has significantly increased with the advances of modern machine learning algorithms. Hidden from the average user's eyes, social media platform providers deploy such algorithms to maximize user engagement through personalized content in order to increase ad revenue. These algorithms analyze users' content, profile users, filter and rank content shown to users. Algorithmic content analysis is also performed by institutions such as banks, insurance companies, businesses and governments. The risk is that institutions attempt to instrumentalize social media with the goal to monitor or even intervene in users' lives [45]. Lastly, algorithms are getting better in mimicking users. So-called social bots [39] are programs that operate social media accounts to post and share unverified or fake content. This includes that modern machine learning algorithms can be used to modify or even fabricate content (e.g., *deep fakes* [53]).

Assessing the risks of these threats is arbitrarily difficult. Firstly, the average social media user is not aware of or vastly underestimates the power of state-of-the-art machine learning algorithms. Without transparency, users do not know what personal attributes are collected, or what content is shown – or not shown – and why. **The lack of awareness and knowledge about machine learning algorithms on the users' part creates an information asymmetry** that makes it difficult for users to effectively and critically evaluate their social media use. Secondly, the negative consequences of users' social media behaviour are usually not obvious or immediate. Concerns such as dynamic pricing or the denial of goods and services (e.g., China's *Social Credit Score* [62]) are typically the result of a long posting and sharing history. The formation of filter bubbles and their effects on users' views is often a slow process. **This lack of an immediate connection between users' behavior and negative consequences prohibits an intrinsic incentive for users to change their behavior.** Lastly, even users who are aware of applied algorithms and negative consequences are in a disadvantaged position. Compared to users, algorithms deployed by data holders have access to a much larger pool of information and to virtually unlimited computing capacities to analyze this information. **This lack of power leaves users defenseless against the algorithmic threats on social media.**

In this paper, we call for solutions to level the playing field – that is, to counter the *information asymmetry* that data holders have over the average users. To directly rival the algorithmic threats, we argue for utilizing machine learning and data mining techniques (e.g., multimedia and multimodal content analysis, image and video forensics) but in a transparent manner and for the benefit of the users. The goal is to help users to better quantify the risk of threats, and thus to enable them to make more well-informed decisions. Examples include warnings that a post contains sensitive information before submitting, informing that an image or video has been

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414692

tampered with, suggesting alternative content from more credible sources, etc. While algorithms towards (some of) these tasks exists (cf. Section 2), we argue that their applicability in our target context of supporting social media users is limited. For interventions (e.g., notices, warnings, suggestions) to be most effective in guiding users behavior, the interventions need to be content-aware, personalized, in-situ, but also trusted and minimally intrusive.

While covering a wide range of research questions from different fields, we believe that the multimedia research community will play an integral part in this endeavour. To this end, we propose a research agenda highlighting open research questions and challenges with focus on multimedia content analysis and content generation. We organize relevant research questions with respect to three core tasks: (1) *Risk assessment* addresses the questions of When and Why social media users should be informed about potential threats. (2) *Visualization and explanation* techniques need to convert the outcome from machine learning algorithms (e.g., labels, probabilities) to convey potential risks in a comprehensible and relatable manner. (3) *Sanitization and recommendations* aim to help users minimize risks either through sanitizing their own content (e.g., obfuscation of sensitive parts in images and videos) or the recommendation of alternative content (e.g., from trusted and unbiased sources). Lastly, given the user-oriented nature of this research, we also highlight the challenges of evaluating the efficacy of interventions in terms of their influence on social media users' decisions and behavior. To summarize, we make the following contributions:

- We present the first comprehensive survey on the threats posed by machine learning algorithms on social media.
- We propose a novel data-driven scheme for user-centric interventions to help social media users to counteract the algorithmic threats on social media.
- We raise open research questions to tackle the *information asymmetry* between users and data holders due to an imbalanced access to data and machine learning algorithms.

Sections 2 and 3 cover our main contributions, forming the main parts of this paper. We complement these contributions by presenting our current proof-of-concept implementation (Section 4) and outlining related research questions and challenges that are equally important but beyond our main research agenda (Section 5).

2 THREATS & COUNTERMEASURES

Arguably, the most prominent threats of social media are loss of privacy, fake news, and filter bubbles or echo chambers. In this section, we review automated methods that facilitate each threat, and outline existing countermeasures together with their limitations. We use the recent global discourse surrounding the COVID-19 pandemic to help illustrate the relevance of these three threats.

Note that social media has also been associated with a wide range of other threats such as online harassment through bullying, doxing, public shaming, and Internet vigilantism [16]. Social media use has also been shown to be highly addictive, often caused by the fear of missing out (FOMO) [15]. Being constantly up-to-date with other's lives often leads to social comparison, potentially resulting in feelings of jealousy or envy, which in turn can have negative effects on users' self-esteem, self-confidence, and self-worth [102]. However, these types of threats are generally not directly caused by algorithms and are thus beyond the scope of this work.

2.1 Loss of Privacy

Each interaction on social media adds to a user's digital footprint, painting a comprehensive picture about the user's real-world identity and personal attributes. From a privacy perspective, identity and personal attributes are considered highly sensitive information (e.g., age, gender, location, contacts, health, political views) which many users would not reveal beyond their trusted social circles. However, with the average user being unaware of the capabilities of data holders and with often no immediate consequences, users cannot assess their privacy risks from their social media use. During the COVID-19 outbreak, many infected users shared their conditions online. Apart from reported consequences such as harassment – e.g., being blamed for introducing COVID-19 into a community – users might also inadvertently disclose future health issues (with the long-term effects of the infection currently unknown).

Algorithmic threats. Given its value, a plethora of algorithms have been designed to extract or infer personal information from multimedia content about virtually all aspects of a user's self: identity (e.g.,[34, 85]), personality (e.g.,[8, 89]), health (e.g.,[24, 56]), location (e.g.,[36, 72]), religious beliefs (e.g.,[77, 109]), relationships (e.g.,[61, 117]), etc. Space constraints prohibit a more detailed discussion here, but we provide a comprehensive overview of the main types of personal information with a selection of related works in our supplementary material. The key message is that algorithms that put user's privacy at risk are omnipresent. While not all methods have been proposed explicitly with social media content in mind (e.g., face detection [48, 49] and action recognition [29, 100] algorithms which are generally used for security surveillance) they can in principle be applied to content shared on social media.

Existing countermeasures. Data and information privacy has been addressed from technological, legal and societal perspectives. With our focus on using technology to protect users' privacy in social media, we categorize existing approaches as follows:

(1) Policy comprehension aims to help users understand their current privacy settings. A basic approach is to show users their profile and content from the viewpoint of other users or the public [7, 64]. More general solutions propose user interface design elements (color-coding schemes, charts, graphs) to visualize which content can be viewed by others [33, 50, 68, 78, 94].

(2) Policy recommendation techniques suggest privacy settings for newly shared data items. Social context-based methods assign the same privacy settings to similar contacts. Basic approaches consider only the neighbor network of a user [4, 32]), while more advanced methods also utilize profile information [5, 38, 51, 69, 71, 92]). Content-based policy recommendations assign privacy settings based on the similarity of data items. First works used (semi-)structured content such as tags or labels [55, 66, 82, 106]. With the advances in machine learning, solutions have been proposed that directly analyze unstructured content with emphasis on images [26, 91, 93, 114] as well as textual content [21, 25, 75]. State-of-the art deep learning models allow for end-to-end learning and yield superior results [90, 112, 113]. More recent policy recommendations

aim to resolve conflicts in case of different privacy preferences for co-owned objects [95].

(3) Privacy nudging [3, 107] introduces design elements or makes small changes to the user interface to remind users of potential consequences before posting content and to rethink their decisions: timer nudges delay the submission of a post; sentiment nudges warn users that their post might be viewed as negative; audience nudges show a random subset of contacts to remind users of who will be able to view a post.

Most of the proposed solutions focus on users' privacy amongst their social circles – that is, privacy is preserved if a user's content can only be viewed by others in line with the user's intention. This assumes that social media platform providers are trusted to adhere to all privacy settings but also that no threats are coming from the platform providers and any data holders with access to users' content. Only privacy nudging does not explicitly rely on trust in data holders. However, current privacy nudges are content-agnostic. They neither provide detailed information why the content might be too sensitive to share nor offer suggestions to users on how to edit or modify content submissions to lower their level of sensitivity.

2.2 Fake News

Mis- and disinformation has long been used to shape people's thoughts and behavior, but social media has significantly amplified its adverse effects. Fake news often leverages users' cognitive biases (e.g., confirmation bias, familiarity bias, anchoring bias), making it more likely for users to fall for it [79]. Fake news is typically also novel, controversial, emotionally charged, or partisan, making it more "interesting" and hence more likely to be shared [104]. During the COVID-19 crisis, misleading claims about infection statistics have resulted in delayed or ignored counteractions (e.g., social distancing measures). False conspiracy theories about the causes for the disease have, for example, resulted in destroying 5G communication towers. More tragically, fake news about supposed cures have already cost the lives of several hundred people.

Algorithmic threats. The most popular algorithmic threat for spreading fake news is the use of bots to spread mis-/disinformation. Particularly bots created with malicious intent aim to mimic humanlike behavior to avoid detection efforts and trick genuine users into following them. This enables the bots to capture large audiences, making it easier to spread fake news. Mimicking humanlike behavior may include varying sharing frequency and schedule, or periodically updating profile information [39, 101]. Apart from better "blending in", sophisticated bots also coordinate attacks through the synchronization of whole bot networks [44]. Recent advances in machine learning also allow for the automated doctoring or fabrication of content. This includes the manipulation of multimedia content such as text swapping [53] or image splicing [30]. When coupled with, algorithms used to detect "infectious" multimedia i.e., content that is most likely to go viral [35, 46, 97], they can be used to predict the effectiveness of fabrication, e.g., for the generation of clickbait headlines [88]. Finally, fake content can also be generated using Generative Adversarial Networks (GANs), a deep learning model for the automated generation of (almost) natural text, images or videos. The most popular example audio-visual content are so-called "deep fakes" [53]: videos that show, e.g., a politician making a statement that never occurred. For textual content, the Generative Pre-trained Transformer 3 (GPT-3) [19] represents the current state of the art of generating humanlike text.

Existing countermeasures. The effects of fake news saw many countries introduce laws imposing fines for its publication [83]. However, the vague nature of fake news makes it very difficult to put it into a legal framework [54] and raises concerns regarding censorship and misuse [104]. Other efforts include public information campaigns or new school curricula that aim to improve critical thinking skills and media literacy. However, these are either one-time or long-term efforts with uncertain outcomes [59]. From a technological perspective, a plethora of data-driven methods have been proposed for the identification of social bots, automated credibility assessment, and fact-checking [87]. Most methods to identify social bots use supervised machine learning by leveraging on the user, content, social network, temporal features, etc. (e.g., [58, 110]). Unsupervised methods aim to detect social bots by finding accounts that share strong similarities with respect to social network and (coordinated) posting/sharing behavior (e.g., [23, 67]). Fact-checking is a second corner stone to counter fake news. However, manual fact-checking – done by websites such as Snopes or Politifact, or dedicated staff of social media platform providers – scales poorly with the amount of online information. Various solutions for automated fact-checking have been proposed (e.g., [47, 52]). However, fully automated fact-checking systems are far from mature and most real-world solutions take a hybrid approach [43].

Existing technological solutions to combat fake news focus on the "bad guys" and do not address the impact of the average user on its success. However, Vosoughi et al. [104] have shown that false information spreads fast and wide even without the contributions by social bots. To evaluate the effects of user nudging, Nekmat [76] conducted a series of user surveys to evaluate the effectiveness of fact-check alerts (e.g., the reputation of a news source). The results show that such alerts trigger users' skepticism, thus lowering the likelihood of sharing information from questionable sources. Similarly, Yaqub et al. [111] carried out an online study to investigate the effects of different credibility indicators (fact checkers, mainstream media, public opinion, AI) on sharing. The effects differ not only across indicators – with fact checkers having the most effect – but also across demographics, social media use, and political leanings.

2.3 Filter Bubble & Echo Chambers

Personalized content is one of the main approaches social media platform providers use to maximize user engagement: users are more likely to consume content that is aligned with their beliefs, opinions and attitudes. This selective exposure has led to the rise of phenomena such as echo chambers and filter bubbles [9, 11, 18]. The negative effects are similar to the ones of fake news. Here, not (necessarily) false but one-sided or skewed information leads to uninformed decisions particularly in politics [13, 37]. Filter bubbles and echo chambers also amplify the impact of fake news since "trapped" users are less likely be confronted with facts or different opinions; as was also the case for COVID-19 [27].

Algorithmic threats. Maximizing user engagement through customized content is closely related to the threats of privacy loss and fake news – that is, the utilization of personal information

and the emphasis on viral content. A such, most of the algorithmic threats also apply here. An additional class of algorithms that often result in "over-customization" are recommender systems based on collaborative filtering [14, 17, 80]. They are fundamental building blocks of many online services such as online shopping portals, music or video streaming sites, product and service review sites, online news sites, etc. Recommender systems aim to predict users' preferences and recommend items (products, songs, movies, articles, etc.) that users are likely to find interesting. Social network and social media platforms in particular expand on this approach by also incorporating the information about a user's connections into the recommendation process [96, 118]. Recommender systems continue to be a very active field of research [12, 116].

Existing countermeasures. Compared to fake news, filter bubble and echo chambers are only a side effect of personalized content [20]. While recommendation algorithms for incorporating diversity have been proposed (e.g., [65, 98]), they are generally not applied by platform providers since they counter the goal of maximizing user engagement. As a result, most efforts to combat filter bubbles aim to raise users' awareness and give them more control [18]. From an end user perspective, multiple solutions propose browser extensions that analyze users' information consumption and provide information about their reading or searching behavior and biases; e.g., [73**?**]. All these approaches can be categorized as nudging by making users be aware of their biases. Despite the numerous nudging measures proposed, their application and acceptance among social media users remain low [60, 108]. This may be because of several reasons. Firstly, as an emerging new technology, digital nudging is not as established as expert nudging, thus may lack users' trust in the first place [108]. Secondly, many people use social media due to its convenience and pleasance, whereas nudging requires extra efforts and attention. Finally, psychologists have found that nudging might only have effects on things people are truly aware of and care about [40].

3 USER NUDGING IN SOCIAL MEDIA

To address the information asymmetry between users and data holders, we motivate automated and data-driven user nudging as a form of educational intervention. In a nutshell, nudges aim to inform, warn or guide users towards a more responsible behavior on social media to minimize the risk of potential threats such as the loss of privacy or the influence of fake news. In this section, we propose a research agenda for the design and application of effective user nudges in social media.

3.1 Design Goals

Effective nudges should be helpful to users without being annoying. Presenting users too often with bad or unnecessary information or warnings may result in users ignoring nudges. We formulate the following design goals for effective nudging:

(1) Content-aware. Warning messages should only be displayed if necessary, i.e., if a potential risk has been identified. For example, an image showing a generic landscape is generally less harmful compared to an image containing nudity. Thus, the latter would more likely trigger a nudge. Similarly, only content from biased or untrusted sources, or content that show signs of being tampered with should result in the display of warnings. Effective nudges need to be tailored to content such as articles or posts being shared.

(2) User-dependent. The need and the instance of a nudge should depend on the individual users, with the same content potentially triggering different or no nudges for different users. For example, a doctor revealing her location in a hospital is arguably less sensitive compared to other users. Similarly, a user who is consciously and purposefully reading articles from different news sites, does not need a warning about each site's biases.

(3) Self-learning. As consequence of both content- and user-dependency, a nudging engine should adapt to a user's social media use. To be in line with the idea of soft paternalism, users should be in control of nudges through manual or (semi-)automated personalization. This personalization might be done through explicit feedback by the user or through implicit feedback derived from the user's past behavior (e.g., the ignoring of certain nudges).

(4) Proactive. While deleting content shortly after posting might stop it from being seen by other users, it is arguably still stored on the platform and available for analysis. Assuming untrusted data holders, any interaction on social media has to be considered as permanent. Thus, nudges need to be displayed before any damage might be done, e.g., before privacy-sensitive content is submitted, a fake news or biased article is shared or even read, etc.

(5) In-situ. Educational interventions are most effective when given at the right time and the right place. Therefore, nudges should be as tightly integrated into users every-day social media use as possible. Ideally, a false fact is highlighted in a news article, sensitive information is marked within an image, warning messages are displayed near the content, etc. The level of integration depends on the environment, with desktop browsers being more flexible compared to closed mobile apps (cf. Section 5).

(6) Transparent. Data-driven user nudging relies in many cases on similar algorithms as potential attackers (e.g., to identify privacy-sensitive information). In contrast to the hidden algorithms of data holders, the process of user nudging therefore needs to be as open and transparent as possible to establish users' trust. Transparency also requires explainability – that is, users need to be able to comprehend why a nudge has been triggered to better adjust their social media use in the future.

3.2 Core Tasks

We argue that an effective nudging engine contains three core components for the tasks of risk assessment, representation and visualization of nudges, and the recommendation or sanitization of content. In the following section, we outline the challenges involved and derive explicit research questions for each task.

Risk assessment refers to the task of identifying the need for nudges in case of, e.g., privacy-sensitive content, tampered content, fakes news or biased sources, clickbait headlines, social bots, etc. As such, risk assessment can leverage on existing methods outlined in Section 2. Note that user nudging therefore relies on the same algorithms used by attackers; we discuss ethical questions and concerns in Section 5. Despite the availability of such algorithms, their applicability for risk assessment is arguably limited with respect to our outlined design goals. Not always is an image containing nudity privacy-sensitive (e.g., an art exhibition), not every bot has a

malicious intent (e.g., weather or traffic bots), not always is a deep fake video shared to deceive but only to entertain users. Effective risk assessment therefore requires a much deeper understanding of content and context. We formulate these challenges with the following research questions:

- How can existing countermeasures against threats in social media be utilized for risk assessment towards user nudging?
- What are the shortcomings of existing methods that limit their applicability for risk assessment with respect to the design goals (particularly to minimize the number of nudges)?
- How to design novel algorithms for effective risk assessment with a deep(er) semantic understanding of the content?

Generation and visualization addresses the task of presenting nudges to users. Sophisticated solutions of risk assessment rely on modern machine learning methods that return results beyond the understanding of the average social media user. Firstly, the outcomes of those algorithms are generally not intuitive: class labels, probabilities, scores, weights, bounding boxes, heatmaps, etc., making it difficult for most users to interpret those outcomes. And secondly, the complexity of most methods makes it difficult to comprehend how or why a method returned a certain outcome. Such explanations, however, would greatly improve the trust in outcomes and thus nudges. The need for understanding the inner workings of machine learning methods spurred the research field of eXplainable AI (XAI) to make such models (more) interpretable. However techniques so far are targeted primarily for experts and improving their usability for end users is an active area of research [1, 2, 41]. Regarding the generation of nudges, we formulate the following research questions:

- To what extent are current XAI methods applicable for user nudging in social media?
- How to convert outcomes and existing explanations into a more readable and user-friendly format for nudging (e.g., charts, content markup or highlighting, verbalization)?
- How to measure the efficacy of nudges along human factors such as plausibility, simplicity, relatability etc. to evaluate the trade-off between these opposing goals?
- How to make the generation of nudges customizable to accommodate users' preferences and expertise or knowledge?

With solutions for generating nudges available, the last step concerns the questions of how to display nudges. Very few works have investigated the effects of different aspects of nudges in the context of privacy [10, 42, 84] and credibility of news articles [63, 76, 81, 115]. Based on previous works, we can define four key aspects for visualizing user nudges: (1) timing, i.e., the exact time when a user nudge is presented, (2) location, i.e., the places where a nudge is presented, (3) format, i.e., the media formats in which the information is presented (whether audio and visual information should be combined with textual information, the length of the information), and (4) authorization, i.e., users' control over the nudging information. We formulate the following research questions for these aspects as follows:

- Given the different threats in social media, what kind of information would users find most useful (e.g., visualization method, level of detail, auxiliary information)?
- How does the *How, When* and *Where* of users' control effect the effectiveness of nudges on the behavior of users?

Recommendation and sanitization expand on nudges that assess and visualize risks of threats to also include suggestions to lower those risks to further support and educate users. In case of fake news, biased sources or tampered content, such suggestions would include the recommendation of credible and unbiased sources, or links to the original content. Regarding our design goals (here: content-awareness) recommended alternative content must be similar or relevant to the original content. This refers to the fundamental task of measuring multimedia content similarity and related tasks such as reverse image search. However, besides the similarity of content, suitable metrics also need to incorporate new aspects such as a classification of the source (e.g., credibility, biases, intention). For recommending alternative content, we propose the following research questions:

- How can existing similarity measures and content linking techniques be applied to suggest alternative content or sources for nudging to lower users' risks?
- How can those methods be extended to consider additional aspects beyond raw content similarity?

In case of users creating content, a more interesting form of suggestion involves the modification of the content to reduce any risks. This is particularly relevant for privacy risks where already minor modifications may avoid a harmful disclosure. However, quickly and effectively editing images or videos is beyond the skills of most users. Content sanitization, the automated removal of sensitive information from content, is a well-established task in the context of data publishing to facilitate downstream analysis tasks of user information without infringing on their privacy. However, with very few exceptions (e.g., [74, 86]), the sanitized content is not intended to be viewed by users. This makes techniques such as word removal, as well as the cropping, redaction or blurring of images or videos valid approaches. In contrast, content sanitization for social media, where the output is seen by others, must fulfil two requirements: *(1) Preservation of integrity.* Any sanitization must preserve the integrity of the content – that is, sanitized content must read or look natural and organic, and it should not be obvious that the content has been modified. *(2) Preservation of intention.* The sanitized content should reflect the user's original intention for posting as much as possible to make it a more valid alternative for the user to consider – formulated as research questions:

- What are limitations of existing content sanitization techniques w.r.t their applicability for nudging in social media?
- How can the integrity of content be measured to evaluate if a sanitized text, image or video appears natural and organic?
- How can a user's intention be estimated to guide the sanitization process towards acceptable alternatives?
- How to design, implement and evaluate novel techniques for sanitizing text, images and videos that preserve both content integrity and user intention?

3.3 Nudging Engine

The nudging engine refers to the framework integrating the solutions for the core tasks of risk assessment, generation and visualization, content recommendation and sanitization, as well as additional task for the personalization and configuration for users.

Frontend. The frontend facilitates two main tasks. Firstly, it displays nudges to the users. For our current prototype, we use a browser extension that directly injects nudges into the website of social media platforms. This includes that the extension can intercept requests to analyze new post before they are submitted; see Section 4 for more details. And secondly, the frontend has to enable the configuration and personalization of the nudging. To this end, the frontend needs to provide a user interface for manual configuration and providing feedback. To improve transparency, configuration should include privacy settings. For example, a user should be able to select whether a new post is analyzed on its own or in combination with the user's posting history. On the other hand, the frontend should also support automated means to infer users' behavior and preferences. For example, if the same or similar nudges gets repeatedly ignored, the platform may no longer display such nudges in the future. The following research questions summarize the challenges for developing the frontend:

- How can nudges be integrated into different environments, mainly desktop browsers and mobile devices?
- What means for configuring and providing feedback offer the best benefits and transparency for users?
- What kind of data should users provide that best reflect their preferences regarding their consideration of nudges?

Backend. The backend features all algorithms for content analysis (risk assessment), content linking (recommendations) and content generation/modification (sanitization). Many of these tasks may rely on external data sources for bot detection, fact-checking, credibility and bias analysis. Depending on users' preferences (see above), the backend will also have to store user content (e.g., users' posting history) and perform behavior analysis to personalize nudges for the individual users. By default, the backend should only keep as much user data as needed – formulated as research questions:

- What are suitable methods to infer user preferences for an automated personalization of nudges?
- Where should user data be maintained to optimize the trade-off between user privacy and the effectiveness of nudges?
- How to identify and utilize relevant external knowledge to complement user data to further improve nudging?

3.4 Evaluation

User nudging as a form of educational intervention is very subjective with short-term and long-term effects on users' behavior. Few existing works have conducted user studies (e.g., for privacy nudges [3, 107]) or included user surveys (e.g., for fake news nudges [76]) in their evaluation. However, evaluating the long-term effectiveness of user nudges on a large scale is an open challenge. While solutions for the core tasks (cf. Section 3.2) can generally be evaluated individually, evaluating the overall performance of the nudging engine is not obvious. We draw from existing efforts towards the evaluation of information advisors. One of the earliest study [99] evaluated a trust-based advisor on the Internet, and used users' feedback in interviews as a criteria for the advisor's efficacy. An effective assessment should serve multiple purposes, measure multiple outcomes, and draw from multiple data sources and use multiple methods of measurements [31].

Similarly, we propose four dimensions from multiple disciplines to evaluate the efficacy of user nudging, namely influence, trust, usage, and benefits. As the most important dimension, *influence* qualitatively measures if and how user behaviour is influenced by user nudging. The *trust* dimension evaluates the confidence of the user in nudges. Since nudges are also data-driven, it is not obvious why users should trust those algorithms more than those of Facebook, Twitter, and the like. *Usage*, as the most objective dimension, measures the frequency of the use of nudges by users. In contrast, the *benefits* for a user are highly subjective as it requires to evaluate how the user has benefited from nudges, which can mostly be achieved through surveys or interviews. Objective and subjective measures from psychology and economics are needed to evaluate the efficacy of user nudging in a comprehensive manner. We formulate the following research questions:

- What are important objective and subjective metrics that quantify the efficacy of user nudging?
- How can particularly the long-term effects of nudges on users' social media use be evaluated?
- How can the effects of nudges on the threats such as fake news or echo chambers be evaluated on a large scale?
- How to evaluate the efficacy of nudging on a community level instead of from an individual user's perspective?

4 SHAREAWARE PLATFORM

To make our goals and challenges towards automated user nudging in social media more tangible, this sections presents ShareAware our early-stage prototype for such a platform.

4.1 Overview to Prototype

For a seamless integration of nudges into users' social media use, we implemented the frontend of ShareAware as a browser extension. This extension intercepts user actions (e.g., posting of new content or sharing of existing content), sends content to the backend for analysis, and displays the results in form of warning messages. The content analysis in the backend is currently limited to basic features to assess the feasibility and challenges of such an approach. In the following examples, we focus on the use case where a user wants to submit a new tweet. If the analysis of a tweet results in nudges, the user can cancel the submission, submit the tweet "as is" immediately, or let a countdown run out for an automated submission (similar to a timer nudge [107]).

ShareAware for privacy. To identify privacy-sensitive information in text, ShareAware currently utilizes basic methods such pattern matching (e.g., for identifying phone number, credit card numbers, email addresses), public knowledge graphs such as WordNet [70] and Wikidata [105] (e.g., to associate "headache" with the sensitive topic of health), and Named Entity Recognition (NER) to identify person and location names. The analysis of images is currently limited to the detection of faces. Figure 1 provides an example. More reliable and effective methods to identify privacy-sensitive information will require a much deeper semantic understanding of

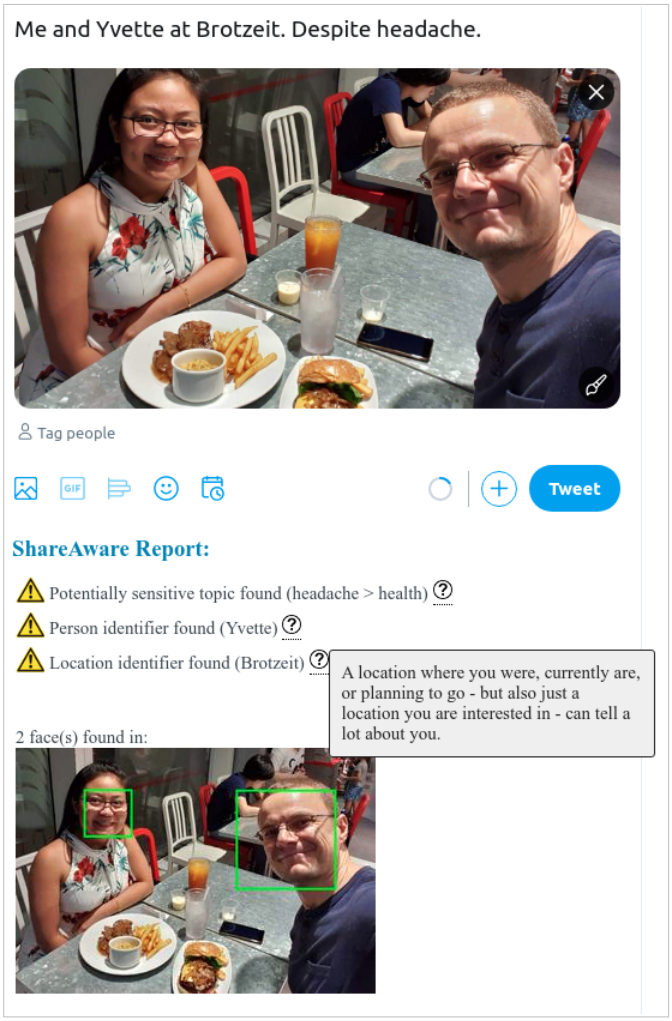

Figure 1: Example of warning messages privacy protection.

shared content. A first follow-up step will be to evaluate algorithms outlined in Section 2.1 for their applicability in ShareAware.

ShareAware against fake news. To slow down the spread of fake news, we display three types of warning messages; see Figure 2. Firstly, we leverage on the Botometer API [110] returning a score representing the likelihood that a Twitter account is a bot. We adopt this score but color-code it for visualization. Secondly, we display credibility information for linked content using collected data for 2.7k+ online news sites provided by Media Bias Fact Check (MBFC).[1] MBFC assigns each news site one of six factuality labels and one of nine bias or category labels. We show these labels as part of warning messages. Lastly, we perform a linguistic analysis to identify if a post reflects the opinion of the tweet author or whether the author refers another source making the statement (e.g., *"Miller said that..."*). In case of the latter, we nudge a user accordingly and ask if s/he trust the source. We present ShareAware for fake news together with an evaluation in a related paper [103].

[1]https://mediabiasfactcheck.com/

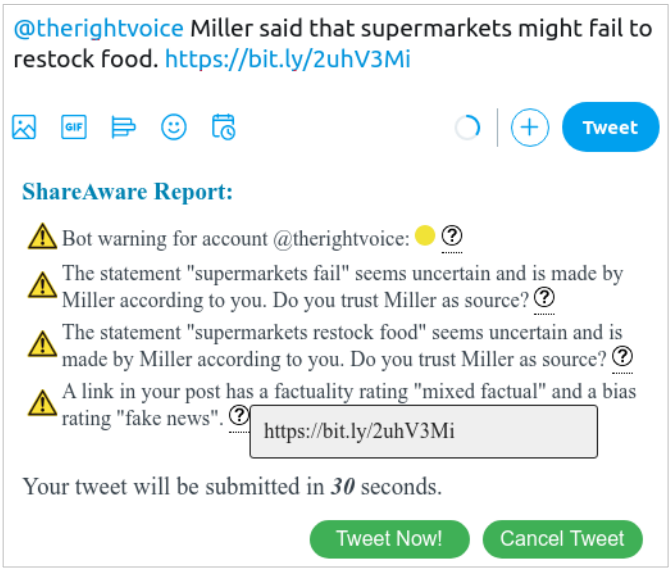

Figure 2: Example of warning messages regarding fake news.

4.2 Experts Feedback

Our current prototype is still in a conceptual stage. To further assess the validity of the concept as well as understand how and if it can be effectively translated into practice, we perform a think aloud feedback session with 4 experts (2 Multimedia + 2 Human Computer Interaction researchers). Each session lasted 45-60 minutes and consisted of discussions centred around different scenarios of use of ShareAware. The sessions were transcribed, and a thematic analysis was performed by 2 members of the research team. While we uncovered many themes, some on implementation and interface design details, we focus on the broad conceptual challenges here.

Explanation specificity. Our experts mentioned that users might find the current information provided by ShareAware rather vague. For example, when ShareAware identified potential disclosure of health related information, they anticipated that users would want to know why this would be of concern to them (e.g., higher insurance premiums). Similarly, in the case of fake news, our experts found the bot score indicators and credibility information insufficient to warrant a change in users' sharing habits. They expected that users would want to see what was specifically wrong about the bot (e.g., spreading of fake news or hate speech).

Explanation depth. ShareAware currently intercepts each post to provide information about potential unintended privacy disclosures individually. However, institutional privacy risks often stem from analyzing historical social media data. Our experts expected ShareAware to provide explanations that help users understand how the current sharing instance would add to unintended inferences made from their posting history. They remarked that a combination of both depth and specificity was required to help users reflect meaningfully and that users should have the option to check the deeper explanations on demand.

Leveraging social and HCI theories. Our experts also felt that for the ShareAware to be effective it had to be grounded in social theories. For example ShareAware could leverage various social theory-based interpretations of self- disclosure in tailoring effecting

nudges or craft nudges which target specific cognitive biases to better help users understand the privacy risks [57]. Similarly, in the example of preventing the sharing of fake news, the interface could incorporate explanations that appeal to social pressure or present posts that share an alternate point of view etc. [22]

Designing for trust. While ShareAware helps users through nudging, it in itself poses a risk due to its access to user data for generating nudges and explanations. Instead of deferring the question of trust to external oversight, legislation or practices such as open sourcing the design, experts felt that ShareAware needed to primarily perform all the inferences on the client side and borrow from progress in areas such as Secure Multiparty Computations [28].

Summing up, the feedback from our experts regarding the current shortcomings of our prototype and future challenges closely match our proposed research agenda. A novel perspective stems from the consideration of social and HCI theories to complement the more technical research questions proposed in this paper.

5 DISCUSSION

This section covers (mainly non-technical) related research questions that are important but not part of our main research agenda.

Ethical concerns. The line between nudging and manipulating can be very blurred. Efforts to guide users' behavior in a certain direction automatically raises ethical questions [3]. Nudging is motivated to be in the interest of users, but it is not obvious if the design decisions behind nudges and users' interests are always aligned. Even nudging in truly good faith may have negative consequences. For example, social media has been used to identify users with suicidal tendencies. Using privacy nudges to help users hide their emotional and psychological state would prevent such potentially life-saving efforts. Nudges might also have the opposite effects. Users might feel belittled by constant interventions such as warning messages. This, in turn, might make users more "rebellious" and actually increase their risky sharing behavior [6]. When, how often and how strongly to nudge are research questions need to be answered before the use of nudging in real-world platforms.

Principle & practical limitations. Algorithms for user nudging based on machine learning generally yield better results when more data is available. As such, social media platforms will likely always have the edge over solutions like ShareAware. Furthermore, a seamless integration of nudges in all kinds of environments is not straightforward. Our current browser extension-based approach is the most intuitive method. On mobile devices, a seamless integration would require standalone applications that mimic the features of official platforms apps, extended by user nudges. This approach is possible for platforms such as Facebook, Twitter or Instagram provide APIs that allow for the development of 3rd-party clients. "Closed" apps such as WhatsApp make a seamless integration impossible. Practical workarounds require the development of apps to which, e.g., WhatsApp messages can be sent for analysis.

Nudging outside social media. In this paper, we focused on data-driven user nudging in social media. However, our in-situ approach using a browser extension makes ShareAware directly applicable to all online platforms. For example, we can inject the user nudges into any website, including Web search result pages, online newspapers, online forums, etc. However, such a more platform-agnostic solution poses additional challenges towards good UX/UI design to enable a helpful but also smooth user experience.

Beyond user nudging. Automated user nudging is a promising approach to empower users to better face the threats on social media. However, user nudging is unlikely to be the ultimate solution but part of a wide range of existing and future efforts. From a holistic perspective of tackling social media threats, we argue that the underlying methods and algorithms required for effective user nudging – particularly for risk assessment and visualization/-explanation – are of much broader value. Their outputs can inform policy and decision makers, support automated, human or hybrid fact-check efforts, improve privacy-preserving data publishing, and guide the definition of legal frameworks.

6 CONCLUSIONS

This paper set out to accomplish two main tasks: outlining algorithmic threats when using social media, and proposing a research agenda to help users better understand and control threats on social media platforms through data-driven nudging. The fundamental goal of user nudging is to empower users to tackle the threats posed due to the information asymmetry between users and platforms providers or data holders. To scale and to compete with these algorithmic threats, user nudging must necessarily use algorithmic solutions, albeit, in an open and transparent manner. While this is multidisciplinary effort, we argue that the multimedia research community has to be a major driver towards leveling the playing field for the average social media user. To kick-start this endeavour, our research agenda formulates a series of research questions organized according to the main challenges and core tasks.

Acknowledgements. This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI research Agenda. In *CHI '20*.
[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *CHI '20*.
[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv.* 50, 3 (2017).
[4] Fabeah Adu-Oppong, Casey K Gardiner, Apu Kapadia, and Patrick P Tsang. 2008. Social circles: Tackling privacy in social networks. In *SOUPS '-8*. Citeseer.
[5] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. ReGroup: Interactive Machine Learning for On-Demand Group Creation in Social Networks. ACM.
[6] M. Amon, R. Hasan, K. Hugenberg, B. I. Bertenthal, and A. Kapadia. 2020. Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE.
[7] Mohd Anwar and Philip W. L. Fong. 2012. A Visualization Tool for Evaluating Access Control Policies in Facebook-style Social Network Systems. (2012).
[8] Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 Personality Traits from Digital Footprints on Social Media: A Meta-Analysis. *Personality and Individual Differences* 124 (2018).

[9] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science* 348, 6239 (2015).

[10] Rebecca Balebako, Florian Schaub, Idris Adjerid, Alessandro Acquisti, and Lorrie Cranor. 2015. The Impact of Timing on the Salience of Smartphone App Privacy Notices. In *ACM CCS '15 Workshops*.

[11] Pablo Barbera, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science* 26, 10 (2015).

[12] Zeynep Batmaz, Ali Yürekli, Alper Bilge, and Cihan Kaleli. 2018. A Review on Deep Learning for Recommender Systems: Challenges and Remedies. *Artificial Intelligence Review* (2018).

[13] Michael A. Beam, Myiah J. Hutchens, and Jay D. Hmielowski. 2018. Facebook News and (De)polarization: Reinforcing Spirals in the 2016 US Election. *Information, Communication & Society* 21, 7 (2018).

[14] Alejandro Bellogin, Ivan Cantador, and Pablo Castells. 2013. A Comparative Study of Heterogeneous Item Recommendations in Social Systems. *Information Sciences* 221 (2013).

[15] David Blackwell, Carrie Leaman, Rose Tramposch, Ciera Osborne, and Miriam Liss. 2017. Extraversion, Neuroticism, Attachment Style and Fear of Missing Out as Predictors of Social Media Use and Addiction. *Personality and Individual Differences* 116 (2017).

[16] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, Article 24 (2017).

[17] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013).

[18] Engin Bozdag and Jeroen Hoven. 2015. Breaking the Filter Bubble: Democracy and Design. *Ethics and Inf. Technol.* 17, 4 (2015).

[19] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[20] Laura Burbach, Patrick Halbach, Martina Ziefle, and André Calero Valdez. 2019. Bubble Trouble: Strategies Against Filter Bubbles in Online Social Networks. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, Vincent G. Duffy (Ed.). Springer.

[21] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. 2014. Privacy Detective: Detecting Private Information and Collective Privacy Behavior in a Large Social Network. In *WPES '14*. ACM.

[22] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction.

[23] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. DeBot: Twitter Bot Detection via Warped Correlation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*.

[24] Lushi Chen, Tao Gong, Michal Kosinski, David Stillwell, and Robert L. Davidson. 2017. Building a Profile of Subjective Well-Being for Social Media Users. *PLOS ONE* 12, 11 (2017).

[25] Lijun Chen, Ming Xu, Xue Yang, Ning Zheng, Yiming Wu, Jian Xu, Tong Qiao, and Hongbin Liu. 2018. A Privacy Settings Prediction Model for Textual Posts on Social Networks. In *Collaborative Computing: Networking, Applications and Worksharing*, Imed Romdhani, Lei Shu, Hara Takahiro, Zhangbing Zhou, Timothy Gordon, and Deze Zeng (Eds.). Springer.

[26] M. De Choudhury, H. Sundaram, Y. Lin, A. John, and D. D. Seligmann. 2009. Connecting content to community in social media via image content, user tags and user communication. In *ICME '09*.

[27] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 Social Media Infodemic. arXiv:cs.SI/2003.05004

[28] Ronald Cramer, Ivan Bjerre Damgård, and Jesper Buus Nielsen. 2015. *Secure Multiparty Computation*. Cambridge University Press.

[29] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. 2019. MARS: Motion-Augmented RGB Stream for Action Recognition. In *CVPR '19*.

[30] Xiaodong Cun and Chi-Man Pun. 2020. Improving the Harmony of the Composite Image by Spatial-Separated Attention Module. *IEEE Transactions on Image Processing (accepted for publication)* (2020).

[31] Joe Cuseo. 2008. Assessing Advisor effectiveness. *Academic advising: A comprehensive handbook* 2 (2008).

[32] George Danezis. 2009. Inferring Privacy Policies for Social Networking Services. In *AISec '09*. ACM.

[33] Ralf De Wolf, Bo Gao, Bettina Berendt, and Jo Pierson. 2015. The Promise of Audience Transparency. Exploring Users' Perceptions and Behaviors Towards Visualizations of Networked Audiences on Facebook. *Telemat. Inf.* 32, 4 (2015).

[34] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[35] A. Deza and D. Parikh. 2015. Understanding Image Virality. In *CVPR '15*.

[36] T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, and N. Deligiannis. 2018. Twitter User Geolocation Using Deep Multiview Learning. In *ICASSP '18*.

[37] Ivan Dylko, Igor Dolgov, William Hoffman, Nicholas Eckhart, Maria Molina, and Omar Aaziz. 2017. The Dark Side of Technology: An experimental Investigation of the Influence of Customizability Technology on Online Political Selective Exposure. *Computers in Human Behavior* 73 (2017).

[38] Lujun Fang and Kristen LeFevre. 2010. Privacy Wizards for Social Networking Sites. In *WWW '10*. ACM.

[39] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. *Commun. ACM* 59, 7 (2016).

[40] Jeff French. 2011. Why Nudging is not Enough. *Journal of Social Marketing* (2011).

[41] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *DSAA '18*. IEEE.

[42] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How Short is too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *SOUPS '16)*.

[43] Lucas Graves. 2018. *Understanding the Promise and Limits of Automated Fact-Checking*. Technical Report.

[44] Christian Grimme, Dennis Assenmacher, and Lena Adam. 2018. Changing Perspectives: Is It Sufficient to Detect Social Bots?. In *Social Computing and Social Media. User Experience and Behavior*, Gabriele Meiselwitz (Ed.). Springer.

[45] S. Gürses and C. Diaz. 2013. Two Tales of Privacy in Online Social Networks. *IEEE Security Privacy* 11, 3 (2013).

[46] Jinyoung Han, Daejin Choi, Jungseock Joo, and Chen-Nee Chuah. 2017. Predicting Popular and Viral Image Cascades in Pinterest.

[47] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster. In *KDD '17*. ACM.

[48] Erik Hjelmås and Boon Kee Low. 2001. Face Detection: A Survey. *Computer vision and image understanding* 83, 3 (2001).

[49] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil K Jain. 2002. Face Detection in Color Images. *IEEE transactions on pattern analysis and machine intelligence* 24, 5 (2002).

[50] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. 2011. Detecting and Resolving Privacy Conflicts for Collaborative Data Sharing in Online Social Networks. In *ACSAC '11*. ACM.

[51] Simon Jones and Eamonn O'Neill. 2010. Feasibility of Structural Network Clustering for Group-based Privacy Control in Social Networks. In *SOUPS '10*. ACM.

[52] Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully Automated Fact Checking Using External Sources. In *RANLP '17*.

[53] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (2020).

[54] David Klein and Joshua Wueller. 2017. Fake News: A Legal Perspective . *Journal of Internet Law* (2017).

[55] Peter Klemperer, Yuan Liang, Michelle Mazurek, Manya Sleeper, Blase Ur, Lujo Bauer, Lorrie Faith Cranor, Nitin Gupta, and Michael Reiter. 2012. Tag, You Can See It!: Using Tags for Access Control in Photo Sharing. In *CHI '12*. ACM.

[56] Enes Kocabey, Mustafa Camurcu, Ferda Ofli, Yusuf Aytar, Javier Marín, and Antonio Torralba andIngmar Weber. 2017. Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media. In *ICWSM '17*. The AAAI Press.

[57] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A Review of Current Research on the Privacy Paradox Phenomenon. *Computers & security* 64 (2017).

[58] Sneha Kudugunta and Emilio Ferrara. 2018. Deep Neural Networks for Bot Detection. *Information Sciences* 467 (2018).

[59] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The Science of Fake News. 359, 6380 (2018).

[60] Matthias Lehner, Oksana Mont, and Eva Heiskanen. 2016. Nudging–A promising Tool for Sustainable Consumption Behaviour? *Journal of Cleaner Production* 134 (2016).

[61] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli. 2017. Dual-Glance Model for Deciphering Social Relationships. In *ICCV '17*.

[62] Fan Liang, Vishnupriya Das, Nadiya Kostyuk, and Muzammil M. Hussain. 2018. Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure. *Policy & Internet* 0, 0 (2018).

[63] Xialing Lin, Patric R Spence, and Kenneth A Lachlan. 2016. Social Media and Credibility Indicators: The Effect of Influence Cues. *Computers in human behavior* 63 (2016).

[64] Heather Richter Lipford, Andrew Besmer, and Jason Watson. 2008. Understanding Privacy Settings in Facebook with an Audience View. In *UPSEC'08*. USENIX

Association.
[65] Gabriel Machado Lunardi. 2019. Representing the Filter Bubble: Towards a Model to Diversification in News. In *Advances in Conceptual Modeling*, Giancarlo Guizzardi, Frederik Gailly, and Rita Suzana Pitangueira Maciel (Eds.). Springer.
[66] Ching Man Au Yeung, Lalana Kagal, Nicholas Gibbins, and Nigel Shadbolt. 2009. Providing Access Control to Online Photo Albums Based on Tags and Linked Data. In *AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0*. Association for the Advancement of Artificial Intelligence.
[67] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. 2019. RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter. In *WebSci '19*. ACM.
[68] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. 2012. The PViz Comprehension Tool for Social Network Privacy Settings. In *SOUPS '12*. ACM.
[69] Julian McAuley and Jure Leskovec. 2012. Learning to Discover Social Circles in Ego Networks. In *NIPS '12*. Curran Associates Inc.
[70] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995).
[71] G. Misra, J. M. Such, and H. Balogun. 2016. IMPROVE - Identifying Minimal PROfile VEctors for Similarity Based Access Control. In *2016 IEEE Trustcom/BigDataSE/ISPA*.
[72] Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying Text, Metadata, and User Network Representations with a Neural Network for Geolocation Prediction. In *ACL '17*. ACL.
[73] Sean A. Munson, Stephanie Y. Lee, and Paul Resnick. 2013. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget. In *ICWSM '13*. The AAAI Press.
[74] M. Murugesan, L. Si, C. Clifton, and W. Jiang. 2009. t-Plausibility: Semantic Preserving Text Sanitization. In *CSE '09*, Vol. 2. IEEE.
[75] Kaweh Djafari Naini, Ismail Sengor Altingovde, Ricardo Kawase, Eelco Herder, and Claudia Niederée. 2015. Analyzing and Predicting Privacy Settings in the Social Web.. In *UMAP '15*. Springer.
[76] Elmie Nekmat. 2020. Nudge Effect of Fact-Check Alerts: Source Influence and Media Skepticism on Sharing of News Misinformation in Social Media. *IEEE Transactions on Dependable and Secure Computing* (2020).
[77] Minh-Thap Nguyen and Ee-Peng Lim. 2014. On Predicting Religion Labels in Microblogging Networks. In *SIGIR '14*. ACM.
[78] Thomas Paul, Martin Stopczynski, Daniel Puscher, Melanie Volkamer, and Thorsten Strufe. 2012. C4PS - Helping Facebookers Manage Their Privacy Settings. In *Social Informatics*, Karl Aberer, Andreas Flache, Wander Jager, Ling Liu, Jie Tang, and Christophe Guéret (Eds.). Springer.
[79] Gordon Pennycook and David G. Rand. 2019. Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking. *Journal of Personality* (03 2019).
[80] Ivens Portugal, Paulo Alencar, and Donald Cowan. 2018. The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review. *Expert Systems with Applications* 97 (2018).
[81] Robin S Poston and Cheri Speier. 2005. Effective Use of Knowledge Management Systems: A Process Model of Content Ratings and Credibility Indicators. *MIS quarterly* (2005).
[82] Ramprasad Ravichandran, Michael Benisch, Patrick Gage Kelley, and Norman M. Sadeh. 2009. Capturing Social Networking Privacy Preferences. In *PETS '09*. Springer.
[83] Peter Roudik. 2019. Initiatives to Counter Fake News in Selected Countries. In *Legal Reports of The Law Library of Congress.*
[84] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A Design Space for Effective Privacy Notices. In *SOUPS '15)*.
[85] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. 2019. Face Recognition Systems under Morphing Attacks: A Survey. *IEEE Access* 7 (2019).
[86] Zhiqi Shen, Shaojing Fan, Yongkang Wong, Tian-Tsong Ng, and Mohan Kankanhalli. 2019. Human-Imperceptible Privacy Protection Against Machines. In *ACM- '19*. ACM.
[87] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (2017).
[88] K. Shu, S. Wang, T. Le, D. Lee, and H. Liu. 2018. Deep Headline Generation for Clickbait Detection. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE.
[89] Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. In *WWW '16*. WWW Steering Committee.
[90] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Adrian Popescu, and Yiannis Kompatsiaris. 2016. Personalized Privacy-aware Image Classification. In *ICMR '16*. ACM, New York, NY, USA.
[91] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2017. Toward Automated Online Photo Privacy. *ACM Trans. Web* 11, 1 (2017).
[92] A. Squicciarini, S. Karumanchi, Dan Lin, and N. DeSisto. 2012. Automatic Social Group Organization and Privacy Management. In *CollaborateCom '12*. IEEE.
[93] A. C. Squicciarini, D. Lin, S. Sundareswaran, and J. Wede. 2015. Privacy Policy Inference of User-Uploaded Images on Content Sharing Sites. *IEEE Transactions on Knowledge and Data Engineering* 27, 1 (2015).
[94] Tziporah Stern and Nanda Kumar. 2014. Improving Privacy Settings Control in Online Social Networks with a Wheel Interface. *J. Assoc. Inf. Sci. Technol.* 65, 3 (2014).
[95] Jose M. Such and Natalia Criado. 2018. Multiparty Privacy in Social Media. *Commun. ACM* 61, 8 (2018).
[96] Jiliang Tang, Xia Hu, and Huan Liu. 2013. Social Recommendation: A Review. *Social Network Analysis and Mining* 3, 4 (2013).
[97] Alexandru-Florin Tatar, Marcelo Dias de Amorim, Serge Fdida, and Panayotis Antoniadis. 2014. A Survey on Predicting the Popularity of Web Content. *Journal of Internet Services and Applications* 5 (2014).
[98] Chun-Hua Tsai and Peter Brusilovsky. 2018. Beyond the Ranked List: User-Driven Exploration and Diversification of Social Recommendation. In *IUI '18*. ACM.
[99] Glen L Urban, Fareena Sultan, and William Qualls. 1999. Design and Evaluation of a Trust-Based Advisor on the Internet. *Interface* (1999).
[100] Gül Varol, Ivan Laptev, and Cordelia Schmid. 2017. Long-Term Temporal Convolutions for Action Recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017).
[101] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization.
[102] Erin A Vogel, Jason P Rose, Lindsay R Roberts, and Katheryn Eckles. 2014. Social Comparison, Social Media, and Self-Esteem. *Psychology of Popular Media Culture* 3, 4 (2014), 206.
[103] Christian von der Weth, Jithin Vachery, and Mohan Kankanhalli. 2020. Nudging Users to Slow down the Spread of Fake News in Social Media. In *IEEE ICME '20 MedFake Workshop*.
[104] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (2018).
[105] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014).
[106] N. Vyas, A. C. Squicciarini, C. Chang, and D. Yao. 2009. Towards automatic privacy management in Web 2.0 with semantic analysis on annotations. In *CollaborateCom '09*.
[107] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. 2014. A Field Trial of Privacy Nudges for Facebook. In *CHI '14*. ACM.
[108] Markus Weinmann, Christoph Schneider, and Jan vom Brocke. 2016. Digital Nudging. *Business & Information Systems Engineering* 58, 6 (2016).
[109] David B. Yaden, Johannes C. Eichstaedt, Margaret L. Kern, Laura K. Smith, Anneke Buffone, David J. Stillwell, Michal Kosinski, Lyle H. Ungar, Martin E. P. Seligman, and H. Andrew Schwartz. 2018. The Language of Religious Affiliation: Social, Emotional, and Cognitive Differences. *Social Psychological and Personality Science* 9, 4 (2018).
[110] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and Generalizable Social Bot Detection through Data Selection. In *AAAI '20*. AAAI Press.
[111] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *CHI '20*. ACM, New York, NY, USA.
[112] J. Yu, Z. Kuang, B. Zhang, W. Zhang, D. Lin, and J. Fan. 2018. Leveraging Content Sensitiveness and User Trustworthiness to Recommend Fine-Grained Privacy Settings for Social Image Sharing. *IEEE Transactions on Information Forensics and Security* 13, 5 (2018).
[113] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan. 2017. iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. *IEEE Transactions on Information Forensics and Security* 12, 5 (2017).
[114] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. 2012. Privacy-aware Image Classification and Search. In *SIGIR '12* (Portland, Oregon, USA). ACM.
[115] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *The Web Conference '18*.
[116] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* 52, 1 (2019).
[117] Yuchen Zhao, Guan Wang, Philip S. Yu, Shaobo Liu, and Simon Zhang. 2013. Inferring Social Roles and Statuses in Social Networks. In *KDD '13*. ACM.
[118] Xujuan Zhou, Yue Xu, Yuefeng Li, Audun Josang, and Clive Cox. 2012. The State-of-the-Art in Personalized Recommender Systems for Social Networking. *Artificial Intelligence Review* 37, 2 (2012).