Towards Understanding and Supporting Journalistic Practices Using Semi-Automated News Discovery Tools

NICHOLAS DIAKOPOULOS, School of Communication, Northwestern University
DANIEL TRIELLI, School of Communication, Northwestern University
GRACE LEE, Medill School of Journalism, Northwestern University

Journalists are routinely challenged with monitoring vast information environments in order to identify what is newsworthy and of interest to report to a wider audience. In a process referred to as *computational news discovery*, alerts and leads based on data-driven algorithmic analysis can orient journalists' attention to events, documents, or anomalous patterns in data that are more likely to be newsworthy. In this paper we prototype one such news discovery tool, *Algorithm Tips*, which we designed to help journalists find newsworthy leads about algorithmic decision-making systems used across all levels of U.S. government. The tool incorporates algorithmic, crowdsourced, and expert evaluations into an integrated interface designed to support users in making editorial decisions about which news leads to pursue. We then present an evaluation of our prototype based on an extended deployment with eight professional journalists. Our findings offer insights into journalistic practices that are enabled and transformed by such news discovery tools, and suggest opportunities for improving computational news discovery tool designs to better support those practices.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**.

Additional Key Words and Phrases: computational journalism; computational news discovery; algorithmic accountability; newsworthiness

ACM Reference Format:
Nicholas Diakopoulos, Daniel Trielli, and Grace Lee. 2021. Towards Understanding and Supporting Journalistic Practices Using Semi-Automated News Discovery Tools. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 406 (October 2021), 30 pages. https://doi.org/10.1145/3479550

1 INTRODUCTION

A lot happens in the world every day—much of it is insignificant, but some of it warrants broader public attention. Journalism, and the news media more broadly, serves to help identify those things happening in the world that may be deserving of wider exposure, scrutiny, or debate in society [74]. This gatekeeping function—of influencing what gets published and publicized as part of the journalistic communication process—is a complex one driven by an array of forces both within and beyond journalism, and at various individual, organizational, social, and technical levels [78]. At the technical level, recent models of gatekeeping have begun to consider algorithmic influences in the context of broader sociotechnical gatekeeping practices [95], including at the information gathering stage of news production [25]. Conceived of as a way to potentially increase the efficiency and scale at which new news stories can be identified [42], such algorithmically informed approaches

Authors' addresses: Nicholas Diakopoulos, nad@northwestern.edu, School of Communication, Northwestern University, Evanston, Illinois; Daniel Trielli, dtrielli@u.northwestern.edu, School of Communication, Northwestern University, Evanston, Illinois; Grace Lee, gracelee@u.northwestern.edu, Medill School of Journalism, Northwestern University, Evanston, Illinois.

2573-0142/2021/10-ART406 $15.00
https://doi.org/10.1145/3479550

can help to orient journalists' attention to events, documents, or anomalous patterns in data that are more likely to be newsworthy [15, 23], a process recently referred to as *computational news discovery* [25, 86].

This work focuses on a particular context of computational news discovery related to the identification of algorithmic decision-making (ADM) systems in government. This is meant to support algorithmic accountability reporting, an emerging beat in journalism focused on investigating and scrutinizing algorithmic decision making across various domains of society [22]. For example, outlets such as *ProPublica*, *Der Spiegel*, and *The Markup* have spearheaded investigative journalism projects seeking to expose and hold accountable power wielded through ADMs.[1] A few efforts have begun to catalogue a range of government use-cases for ADMs in the U.S., from services administration to regulatory enforcement [18], with one study finding that nearly half of the 142 surveyed agencies had used various AI and machine learning tools [31]. While some government initiatives have begun to develop responsible approaches to ADM deployment or even create registries of algorithms in use, such approaches are not widespread [35, 92]. The need for scrutiny of these systems is evident [32, 33] as is the need for additional research on the responsible design, development, use, and evaluation of ADMs deployed throughout the public sector [16, 36, 49, 61, 73, 93]. In this work we focus on developing and evaluating a method and tool to help comprehensively monitor and track ADM usage across all levels of government to support journalism.

An important underlying motivation for the current work is to find ways to reduce the effort needed to engage in the scrutiny of government algorithms. This research therefore undertakes the design and evaluation of a semi-automated computational news discovery tool called *Algorithm Tips* to enable such work. Algorithm Tips systematically and periodically monitors government websites (across U.S. government at all scales) for documents that may reveal new applications of ADMs. Through a series of automated evaluations, internal expert evaluations, and crowdsourced evaluations Algorithm Tips augments documents to produce news leads. These leads are presented in an online interface and sent to external professional journalists who ultimately decide whether to pursue the additional reporting needed to transform a lead into a publishable news story. We evaluated Algorithm Tips in an extended deployment with eight professional journalists who have experience reporting on algorithms in society for some of the largest most well-established news organizations in the U.S. Our findings articulate ways in which the tool meets the needs and practices of domain experts, and offers implications for the design and further research of computational news discovery tools more broadly.

This research offers a couple of contributions, including (1) the design and development of a computational news discovery tool (Algorithm Tips) with design goals tailored to enable journalistic decision making around which leads to pursue in the domain of algorithmic decision making in government; and (2) an evaluation of that tool with eight professional journalists in an eight-week-long deployment which offers ecologically valid insights into how journalistic practices are enabled and transformed by such a news discovery tool. In particular we describe design goals related to supporting attention management, as well as verification and newsworthiness professional evaluations, and we tailor the information process and interface of Algorithm Tips to support those goals in the context of investigative and enterprise reporting. Our findings elaborate on how journalists made context-specific interestingness decisions using the evaluative information made available via automation, crowds, and experts; saw the tool as able to offer both a jumping-off

[1]See: Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. ProPublica, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Blackbox Schufa. Der Spiegel. November, 2018. https://www.spiegel.de/wirtschaft/service/schufa-so-funktioniert-deutschlands-einflussreichste-auskunftei-a-1239214.html; and Swinging the Vote? The Markup. February, 2020. https://themarkup.org/google-the-giant/2020/02/26/wheres-my-email

point for specific items and to provide background on the beat of government algorithms; and strove to manage their own effort and attention with respect to the tool. Overall our work provides insight into how to more effectively design and build computational tools to support the essential journalistic activities of news discovery and news gathering.

2 BACKGROUND

In designing and developing Algorithm Tips we were informed by related work in (1) the domain of Computational Journalism as it relates to Human-Computer Interaction (HCI), including various tools that have addressed the topic of news discovery in different contexts; and (2) the topic of newsworthiness and the measurement of news values as it may support journalistic evaluations of potentially newsworthy information.

2.1 Computational Journalism

Studies of journalism within the context of HCI [3] address a broad gamut of news-related activities, from information gathering [26, 29], production [56, 64, 81], and sensemaking [14, 28], to distribution [80, 89] and audience consumption behavior [8, 34, 63]. Oftentimes such research is oriented towards designing and developing new tools to augment journalists' capabilities, or towards seeking to more broadly understand how journalists use computational tools in their sociotechnical practices [47, 88]. The current work contributes to this growing body of HCI literature on computational journalism, which itself is broadly oriented towards enabling information and knowledge production using algorithmic approaches that embrace journalistic values and advance journalistic practices [19, 23]. In particular we focus on the goal of supporting news discovery and gathering tasks [25, 70], which prior research has documented as challenging and time-consuming for journalists [47, 81].

In designing Algorithm Tips we were informed by prior work on computational news discovery tools. Early ideations about the role of computing in journalism suggested that algorithmic techniques might offer an information subsidy to the discovery and gathering phases of news production by allowing journalists to scan the environment for newsworthy events and happenings at increased scale, speed, and overall efficiency [42, 43]. Recently this has been framed as *computational news discovery* (CND), defined as "the use of algorithms to orient editorial attention to potentially newsworthy events or information prior to publication" [25]. The premise for CND is supported by work in journalism studies showing that (non-computational) sources and information subsidies that are frequently used and routinely relied on do tend to save reporters' time during news discovery [71].

Some of the first CND tools were developed in the context of social media monitoring to help detect newsworthy events, identify witnesses, or make sense of how people were responding to news events [26, 28, 76]. Given the scale and quality of social media data confronting journalists many of these tools support not only discovery tasks [30], but also curation and validation tasks [54, 62, 88, 100] which emphasize the importance of assessing and verifying sources and content [12]. CND tools have also been built to support a variety of use cases beyond social media listening, including the monitoring of numeric data streams or sets (e.g. crime, real estate, education) for trends, outliers, or anomalies [15, 55, 77], the monitoring of local court or business documents to surface newsworthy people, places, or companies [67], identification of local public meeting events that journalists may want to attend [97], and the ranking of claims in the media that may be worthy of fact checking [40, 46].

In this work we draw on this prior work to help elaborate design goals and inform key features for the design of the Algorithm Tips tool (See Section 3.1). We also distinguish the tool in its orientation towards monitoring and generating leads about the use of algorithmic decision making in government, a particular context that has not been addressed in prior work, but which is indicative

of a larger class of investigative reporting efforts to monitor government documents that might be similarly supported technologically. Moreover, the evaluation of many of the tools developed in prior work has been based on case studies, or short-term scenario-based usage, rather than extended deployments (excepting [76]). In contrast, in this work, we report on a two-month-long deployment of Algorithm Tips with professional journalists in an effort to better study the use of such a tool as it relates to more ecologically valid journalistic workflows and practices [79].

2.2 News Values and Newsworthiness

Much of the research on tool development for journalists tends to emphasize the importance of aligning designs with journalistic values [23, 24, 47, 51]. We address this observation in this section by presenting a closer examination of how CND tools might incorporate measurements of news values to support journalistic decisions of newsworthiness. Whether an event or occurrence in the world actually becomes a news story is the result of a sociotechnical gatekeeping process that is influenced by forces at various individual, organizational, social, normative, economic, and technical levels [78]. Yet research has shown that a number of factors, termed *news values*, have been repeatedly observed and are manifest in the types of stories that journalists tend to report and publish. For instance, journalists may consider news values such as timeliness, proximity, prominence, human interest, relevance, conflict / controversy, unexpectedness / surprise, reference to the power elite, audience fit, actuality, and consequence / impact, among others [5, 44, 45, 65, 75]. Whether a news lead is deemed newsworthy enough to become a news story arises out of a contextualized judgement process informed by these news values. In this paper we consider how to integrate measures of news values, or newsworthiness, into news discovery processes and tools in order to *inform* (but not supplant) professionals' editorial judgements. In particular, we consider how different news values may be amenable to measurement using computational and crowd-based techniques (elaborated further in Sections 3.2.2 and 3.2.4), while acknowledging in our interface design that professionals should be enabled to make the final contextual judgment of whether and how a lead fits and is positioned at broader organizational and societal levels of interest [75].

All of the news discovery tools referenced above encode and reflect computational operationalizations of news values in some way, oftentimes in domain- and data-specific ways. For instance, several efforts have tried to capture the dimension of "unexpectedness / surprise" by detecting outliers or anomalies in numerical data streams or sets [27, 55, 77]. The Lead Locator system additionally measures "political relevance" in a domain-specific way, and includes a basic operationalization of "magnitude" based on population sizes, which makes sense in the political reporting context [27]. The Vox Civitas system uses measures of "relevance" and "uniqueness" based on text analysis and cosine similarity to help filter social media posts in response to political speeches [29]. The Local News Engine captures the dimension of "reference to elites" by automatically extracting named entities of people, places, or companies from local documents [67]. The Reuters Tracer system explores several operationalizations of news values for evaluating detected events in social media, including "topical relevance", "scale" (or "magnitude" of the event), "negative impact" including human, physical, and financial impacts, "location", and "novelty" of the event [62]. In some cases news values are based on fairly straightforward statistical measures [27, 55] while in others sophisticated machine learned models [62] or formal semantic models [60] are applied. There have also been a handful of prior efforts to measure news values in non-discovery contexts, such as for evaluating news headlines [68] or for predicting the sharing of news articles socially [91]. Given this prior work, and taking into account the specific context explored in the current work (i.e. focused on administrative documents related to the use of algorithms in government) we felt confident in computationally supporting four news values: timeliness, proximity, reference to the power elite, and topical relevance (See Table 1).

In contrast to prior work on computational news discovery systems, this research also explores the use of crowdsourcing techniques to measure news values, and more importantly, evaluates these crowdsourced signals of newsworthiness with domain experts in a specific and extended deployment context. Crowdsourcing techniques have previously been used throughout journalistic processes for everything from checking documents and rating claims and content for factchecking or verification purposes, to identifying and verifying locations and context, co-developing topics for coverage, helping to moderate and route information in communities, gathering information, and even writing news articles [1, 2, 21, 50, 52, 94, 101]. This work distinguishes itself from prior work by exploring the applicability of directly crowdsourced measures of news values in helping to augment professional judgements of newsworthiness. Crowdsourced rating approaches have been shown to be effective for the reliable measurement of a range of manifest and latent constructs in communications content [7, 28, 53], which we extend here to apply to news values. We consider both quantitative evaluations of news values (i.e. numerical ratings on a 1-5 scale) and qualitative rationale for those ratings as an added dimension of context and information for end-users [57]. In particular we apply this approach to measuring news values that we think would benefit from the diverse social knowledge and independent evaluations that can be captured through crowdsourcing [84], including: bad news (which we frame more specifically as "negative societal impacts"), magnitude of impact, controversy/conflict, and unexpectedness/surprise (See Table 1). These measures are included in the Algorithm Tips user interface and are considered in our evaluation in terms of how journalists incorporate such crowdsourced measures of news values into their broader newsworthiness assessments.

3 ALGORITHM TIPS DESIGN

Here we describe the design of Algorithm Tips both in terms of how information flows through the system and is augmented and evaluated, as well as in terms of how end-user journalists are able to view and interact with that information. The system was initially developed and piloted in early 2017 [90], and since then has gone through several design iterations to simplify the information architecture and design, reduce the amount of manual effort needed while taking advantage of automation where possible, and incorporate a novel crowdsourcing component. These iterations were spurred on by informal feedback from several professional journalists working at established news organizations who were shown early versions of the interactive prototype and methods either individually or in the context of sessions at practitioner conferences or workshops that were geared towards soliciting feedback. For instance, such feedback informed features to support the evaluation of newsworthiness including topical and date-based filtering and the highlighting of key entities in documents.

We designed Algorithm Tips to support investigative or enterprise journalism [69] around the use of algorithms in government. Investigative or enterprise journalism reflects a typical orientation towards algorithmic accountability reporting [22, 23] and is characterized by extended time frames [42] with less of an emphasis on the speed or immediacy of the report with respect to events in the world [99]. Moreover, the specific reporting context of algorithms in society that we seek to support does not entail an exhaustive review of *all* leads. The goal is rather to signal when something is interesting and worthy of attention rather than demand the journalist exhaustively find everything of interest. We focus on supporting the initial news discovery phase (i.e. the initial contact or awareness a journalist receives about a potential lead) and on the transition into the news gathering process [70]. This design orientation has implications for how the tool is configured (e.g. to generate alerts on a weekly basis), what types of journalists might expect to find value from the tool (e.g. investigative or enterprise rather than daily or breaking), and the temporal affordances

of the leads the tool might surface (e.g. oriented towards longer term issues and trends rather than events) [85].

After analyzing the extant computational news discovery systems and tools from the literature (See Section 2.1), and drawing heavily on [25], in the next section we articulate several more specific design goals for Algorithm Tips. The final flow of information through the system, from monitoring documents on the web, to suggesting news leads to professional journalists is depicted in Figure 1. The information process used to harvest and enrich documents is described in Section 3.2 and the interface used to present leads to end-user journalists is described in Section 3.3, both of which are described in terms of how they help support the overall design goals.

3.1 Design Goals

- **DG1-Attention:** *Be sensitive to available human effort and attention.* Journalists are already overwhelmed with information from many different channels, sources, and information subsidies (e.g. press releases) [47, 81]. A lead discovery tool should be sensitive to the journalist's attention economy by aligning suggested leads to their interests [25]. This can be supported through search and filtering that allows the information space to be interactively adapted, alerts that are configurable and schedulable to suit individual interests and timing needs, and support for marking of items so that users can come back to initially interesting leads when more attention is available.
- **DG2-Verification:** *Enable initial verification and quality assessment.* In order for a lead to turn into a substantive piece of journalism it must be verified and confirmed as accurate [88]. This is an iterative process that unfolds as the lead is pursued and new information is collected, however, even at the initial phase a lead discovery tool should enable verification and information quality assessment activities [25]. This can be supported by including adequate context to evaluate or verify information and launch into follow-up research, such as by including a link to the original source document, provenance information on the search terms and search used to find the document, and date/time information to indicate the timeliness of the information [12].
- **DG3-Newsworthiness:** *Enable newsworthiness evaluation.* Journalists apply a range of context-specific criteria to a news lead in determining whether they think it is important enough to pursue and develop into a publishable news story [25]. These may include factors such as timeliness, proximity, significance, novelty, various dimensions of relevance, as well as whether it is likely to be of interest to their imagined audience. Journalistic decisions of newsworthiness can be supported by incorporating relevant metadata in the information design so that journalists can develop their own nuanced interpretations of these various factors.

3.2 From Documents to Leads

In this section we describe the design of the first four stages of the Algorithm Tips information flow, beginning with the document monitoring apparatus. After documents are collected they are then evaluated, augmented, and filtered in several ways before they become leads that are delivered to journalists. In the following subsections we first describe the monitoring method, and then describe the details of different processes for automated, expert, and crowdsourced evaluation of the documents (See Table 1) which help support the design goals described in Section 3.1.

3.2.1 Document Monitoring. A method for targeted Google web searches and scraping was developed in order to automatically monitor for potentially interesting instances of ADM systems

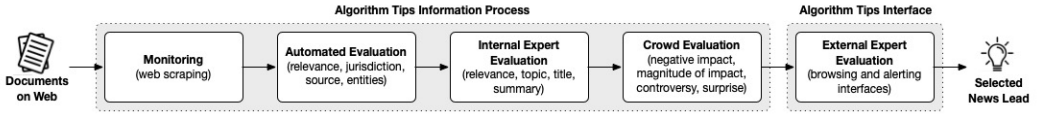


Fig. 1. The Algorithm Tips information process and interface.

used in government. In order to develop a list of search queries to target our concept of interest (i.e. ADMs) we considered how algorithms might be referred to across different government agencies (DG1-Attention). From our experience with public records requests for algorithms from state agencies we knew that the vocabulary used to describe algorithms internally could vary. For instance, an algorithm for predicting criminal recidivism might be referred to as a "risk assessment" by the relevant government agency. We created a list of algorithm-related terms such as "algorithm", "predictive analytics", "risk assessment" and so on. Initially, 61 terms were identified, based on brainstorming and the use of various thesauri. As described in [90] we then used term expansion and redundancy identification techniques to both widen the net and enhance the coverage or recall of the system while also striving for a non-redundant set of queries that would increase efficiency and be considerate of the demands the scraper put on external systems [9]. This process yielded a final set of 67 terms.[2]

Since our initial focus and area of interest is governmental algorithms we limit searches to U.S. government domains. This was done by appending the "site:.gov" operator to the search. Initial results included a high degree of noise due to research papers hosted on government websites that described their methodologies using many of the search terms. A majority of those research papers were hosted by the National Institutes of Health (NIH). While informative, early assessments revealed that many of these references would not necessarily lead to algorithms being used by government agencies. As a pragmatic response, they were therefore filtered out of the searches by using another operator, "-site:.nih.gov".[3] In the end, searches for data collection were constructed using the following format: *"[search term]" site:.gov -site:.nih.gov.*

Searches are automatically executed by our system for all of the search terms tracked on a periodic basis (once per week in the current setup). Up to 100 results for each term are collected although a document (e.g. HTML, text, or PDF file) is excluded if its URL is a duplicate of a document URL found with a previously searched term. Also, if a document URL is found in subsequent rounds of data collection (e.g. in a future week) it is marked as a duplicate and only the first occurrence is retained. In an effort not to overload Google servers we only submit one search query every 1-2 minutes.

3.2.2 Automated Evaluation. Each document that is collected is automatically tagged with a few pieces of metadata to support the evaluation of several news values (DG3-Newsworthiness) (See Table 1). To support evaluations of the news value of "timeliness" we store a date-time stamp for every document, both for when it is initially found online, and for when it is finally published on Algorithm Tips. To augment documents with data about both the relevant jurisdiction (e.g. "City", "State / Local Government", or "Federal Agency — Executive") and the government source that published information about the algorithm (i.e. an agency, organization, or municipality), we use the domain name of the URL where the document was found together with curated data provided by

[2]The final set of terms is available at REDACTED.

[3]Other government domains which might be expected to yield many research papers, such as nsf.gov, only accounted for a small minority of documents (~1%) and were therefore not explicitly filtered out in order to maintain as wide a net as possible.

the U.S. General Services Administration which links jurisdictions and sources to domain names.[4] This information is meant to support the news value of "proximity" by conveying to end users the geographic scope of the lead. In order to support the news value of reference to "the power elite", we automatically extract named entities for all people or organizations mentioned in each document.[5] Based on frequency of occurrence in the document the top five people and the top five organizations are retained.

In addition, we developed a custom model for assessing document relevance which scores each document on a scale from 0 (not relevant) to 1 (relevant) with respect to the topic of algorithmic decision making (DG3-Newsworthiness). This is a measure of topical relevance and is meant to align the editorial scope of Algorithm Tips' leads with the presumed interest of external experts coming to Algorithm Tips for leads (DG1-Attention). To train the model we manually annotated a random subsample of documents collected via our scraping process. Data was collected across three separate dates in 12/2016, 3/2018, and 6/2019 with a goal of contributing to greater resilience of the model to temporal shifts in the data. Because text could not be extracted from all documents and due to link-rot between the collection and annotation, the final sample consisted of 223 relevant documents and 304 irrelevant ones. To ensure annotation reliability we developed guidelines in an annotator guidebook which includes clear definitions and criteria for relevance as well as positive and negative examples. We trained a second coder using the guidebook and on a random sample of 100 documents achieved an inter-rater reliability Cohen's Kappa of 0.87 with one of the co-authors, indicating a reliable coding process. The second coder then annotated the remainder of the documents. The final set of 527 annotated documents was then used to train a model, with features including all term frequency inverse document frequency (TF-IDF) scores of unigrams, bigrams, and trigrams and excluding English stopwords.

To establish the appropriate criteria for the model evaluation we first consider its intended use context. In particular, we assume that journalist end-users will have a limited amount of time available to review leads, and that they will be satisfied with finding some subset of reasonable leads without expecting to comprehensively find *all* leads (DG1-Attention). We similarly assume the internal expert evaluator has a fixed time budget but has the goal of increasing yield (i.e. finding as many relevant leads as possible in the time budget). A high precision will increase the yield. This use context therefore suggests that the classifier should be evaluated using the precision@k metric, which considers the precision of a ranked set of results of size k (where k controls the fixed time budget). We experimented with Bayes, Logistic Regression, Support Vector Machine, and Random Forest algorithms for learning the model and compared their performance with k = 10, 25, and 50. All models were trained using 5-fold cross validation by training on different subsets of 80% of the data to generate predicted scores for the 20% left out.

We found that the Logistic Regression classifier outperformed the others, yielding a P@25=0.6 and P@50=0.6. So, for example, on a sample of 25 documents ranked with the model we would expect 15 of them (i.e. 60%) to be relevant. In comparison, a random sample of 25 would be expected to yield about 10.58 relevant documents (i.e. 223 / 527 or ~42% is the base rate of relevant documents in the sample). Given a fixed time budget that allows for the examination of, for instance, only 25 documents, the model is expected to boost efficiency by yielding about 42% more relevant documents in comparison to a random selection of documents (i.e. from ~10.58 to 15). The model is therefore useful insofar as it should increase the yield of relevant documents identified by the internal expert given their fixed time budget. Alternatively, the model could be deployed to *reduce*

[4]https://github.com/GSA/data/tree/master/dotgov-domains

[5]We use the SpaCy large model for named entity recognition (https://spacy.io/usage/facts-figures) which has an accuracy very close to the state of the art at about 86%

the time budget but maintain a fixed yield, however we opted not to use the model this way as we want to tune the system to find as many relevant documents as possible. Admittedly there is still likely room for improvement here, such as by collecting more training data, undertaking more sophisticated feature engineering, or applying more advanced models. However, because the absolute performance of this model is not the main focus of this paper, we deem this performance to be adequate to our use context, particularly in light of the internal expert evaluation we describe next.

3.2.3 Internal Expert Evaluation. Our document evaluation workflow utilizes expert evaluation at several levels from both internal (project members) and external (recipients of leads) experts (See Figure 1). For our purposes here we consider "experts" to be highly trained individuals with incisive judgement and ability to evaluate documents with respect to journalistic needs. In order to train internal experts we developed a guidebook which walks the internal expert through the process of evaluating document relevance and producing metadata. Internally, expert evaluation is necessary to do quality checks on the automated evaluations (DG2-Verification), and to provide information transformations that we were not able to automate.

Perhaps most critically, internal experts make a final determination of document relevance and are thus the final gatekeeper for whether a document is promoted to being a lead (DG1-Attention; DG3-Newsworthiness). Here the goal is to ensure high precision of the leads that are ultimately delivered to external experts. In order to manage the scale of monitoring and adhere to their time budget the internal expert only makes this judgement for a subsample of documents that are automatically ranked by relevance within each data collection period. This involves carefully reading the document to evaluate whether it really describes an algorithmic decision making system, or whether it was perhaps spuriously considered relevant by the model. In making this judgement we apply a pragmatic definition that captures a broad range of the types of algorithms typically encountered by the system: "An algorithm is a set of rules to which data can be input and from which a result, such as a score, a calculation, or a decision, is obtained. Algorithms can be either computational (e.g., computer software or a spreadsheet) or not computational (e.g., a weighted score card or flowchart that could be applied by a person)."

Internal experts also evaluate the appropriateness of the extracted entities and remove any that are inaccurate (DG2-Verification). They do so by manually reviewing the documents from where those entities were extracted to determine the context of each entity. This allows the internal experts to determine which entities are journalistically relevant, and which can be excluded. Relevant entities are government agencies and programs or companies that are in some way involved with the algorithm. Irrelevant entities are authors of reports or individual government workers, and common words or concepts that are incorrectly extracted by the automated process (e.g. a document related to COVID-19 might have entities such as "CDC" and "PPE" automatically detected; the former is relevant, since it is a government agency, and the latter is not, since in this context it means "personal protective equipment"). These steps are important in order to keep the relevance of leads provided to external experts as high as possible (DG3-Newsworthiness), and with metadata that is as accurate as possible since low accuracy can be a barrier to the use of monitoring tools (DG2-Verification) [48].

The internal expert is also responsible for writing a title and short summary (1-2 sentences) of the ADM described in each document (DG1-Attention). Because these texts are used in further crowdsourcing, they are written without jargon or acronyms and are self-contained explanations of what the ADM is, with details on what it does and who is using it. Finally, the expert also adds a

manually selected topic field to the lead (e.g. "Energy", "Transportation and Public Works") based on a set of 32 policy areas defined by the U.S. Congress (DG3-Newsworthiness).[6]

3.2.4 Crowdsourced Evaluation. Here we describe our method for collecting suitable evaluations from crowdworkers to support journalists in making their newsworthiness assessments (DG3-Newsworthiness). Based on the news values literature cite above we identified four newsworthiness factors that rely on more widespread social knowledge (rather than expert or domain-specific knowledge) and which we therefore thought may be amenable to gathering from crowdworkers. These include (1) Bad News, which we adapt to Negative Societal Impacts, i.e. "The algorithm described has the potential to create negative impacts in society"; (2) Magnitude, which we frame as Number of People Impacted, i.e. "The algorithm described has the potential to involve or impact a substantial number of people"; (3) Potential for Controversy, i.e. "The algorithm described has the potential to be controversial in society"; and (4) Surprising, Unusual, or Unexpected, i.e. "The algorithm described was surprising, unusual, or unexpected". Other news values may also be appropriate to crowdsource, but we believe that starting with these four is a reasonable first step. See Table 1 for definitions stemming from the news values literature.

In line with best practices in crowdsourcing we pay careful attention to task design, working through several iterations to improve the clarity of instructions [4, 57, 98]. The final design of each crowd task provides instructions which define what an algorithm is, and asks the worker to think about how the algorithm would be used and would impact people (See the the Appendix for the full details on the layout and wording of the task). The title, summary, and government source of the document are provided, with a link to the original document which workers may optionally consult. Workers are asked to rate the document on a scale from 1 (completely agree) to 5 (completely disagree) for each statement about each factor, e.g. "The algorithm described has the potential to create negative impacts in society". We chose to frame the task as an agree/disagree format to emphasize the subjectivity of the question and to trigger the raters to assess the document based on their own point of view on the algorithm [37]. In addition, workers are asked to provide a rationale for their rating, which has been shown to increase crowdsourced rating quality and also offers qualitative context for the rating to end-users [57].

As this crowdsourcing task is subjective and interpretive of latent (i.e. non-manifest) constructs we do not compute validity measures such as inter-rater reliability [72]. Instead we rely on aggregation (a simple mean in this case) to capture the central tendency of the ratings, which has shown to yield results that correlate well with expert judgment in other content analysis tasks [7, 53]. The benefit of this approach is that it helps capture the diversity of subjective interpretations about how the algorithm described in each document may be newsworthy from each individual rater's point of view.

In our deployment, which uses Amazon Mechanical Turk, we asked five independent workers to rate each document, offering $0.50 per rating based on the median time taken on a pilot task and a target of paying at least minimum wage in the U.S. Workers were screened to be from the United States in order to to ensure knowledge of cultural context for the evaluation of the government documents. In an effort to enhance quality workers were required to have an approval rate greater than 99% and number of prior approved tasks greater than 500.

3.3 User Interface Design

The user interface of Algorithm Tips supports the last stage of the information flow through the system (See Figure 1). The overarching design goal was to develop an interface that would help professional journalists identify promising leads for further reporting. The following subsections

[6] See: https://www.congress.gov/help/field-values/policy-area

Table 1. News values and the source of evaluation of each in Algorithm Tips.

News Value	Definition	Source
Timeliness	Items that reflect new or current information [75]	Automated; External Expert
Proximity (geographic)	Items more nearby according to geographic distance between event location and publication location [5]	Automated; External Expert
The Power Elite	Items "concerning powerful individuals, organizations, institutions, or corporations" [45]	Automated; Internal Expert; External Expert
Relevance (topical)	Items about issues that are expected to be important to the public or to a specific audience [44, 75]	Automated; Internal Expert; External Expert
Bad News	Items with "particularly negative overtones such as death, injury, defeat and loss" [45]	Crowd; External Expert
Magnitude	Items "perceived as sufficiently significant in the large numbers of people involved or in potential impact, or involving a degree of extreme behaviour or extreme occurrence" [45]	Crowd; External Expert
Controversy / Conflict	Items "concerning conflict such as controversies, arguments, splits, strikes, fights, insurrections and warfare" [45]	Crowd; External Expert
Unexpectedness / Surprise	Items having "an element of surprise, contrast, and/or the unusual about them" [45]	Crowd; External Expert
Actuality	Items having a relation to the present moment and occurrence of events in the world [5]	External Expert
Organizational Agenda	Items that "fit the news organization's own agenda, whether ideological, commercial or as part of a specific campaign" [45]	External Expert

describe how the information and interaction design of the system helps support the design goals described in Section 3.1.

3.3.1 Interface and Interaction Design. The Algorithms Tips interface (See Figure 2) is designed as on online web application comprised of several tabs that the user can switch between, including tabs for viewing leads from the database (DG1, DG2, DG3), configuring alerts (DG1), and reviewing flagged leads (DG1). A help screen provides an explanation and transparency into how the leads are found and augmented by the system (DG2). Users can sign in to the app using a Google login, which allows them to configure and save alerts and flags. After configuration, alerts are sent to users as formatted emails, a modality that prior work suggests is appropriate for notifications relating to information gathering monitoring tasks [48], and which is familiar and convenient to many journalists as a channel for receiving and archiving information digests [55, 81].

The default interface view for Algorithm Tips shows a listing of leads from the database (the information design of these is described in the next section). In order to provide a clear and understandable display, the ranking of leads is simply based on publication date, however different or dynamic rankings is an interesting area for future work as that has been effective in other tools for journalists [64]. Users can then search and filter the leads in order to refine the leads displayed

according to their particular interests (DG1). The implementation of search matches search terms across the title, summary, topic, and entities fields. Filtering can be applied based on the publication date or based on the source of the lead, such that users can tailor their interests to, for instance, state or local sources of documents published in the last week.

The Alerts interface displays a listing of saved alerts for the user. Users can edit or delete alerts once created. When creating an alert the user sets the search terms, source filter, frequency, and recipient, thus facilitating some control over the tailoring and scoping of the information that they will be alerted to (DG1) [48]. Before saving, alerts can be previewed against existing leads in the database to help users verify that the configuration is yielding what they would expect. The frequency of alerts can be set for weekly, semi-weekly (every 10 days), or monthly so that users can somewhat control how often they receive emails from the system (DG1).

The user interface for the alerting system in Algorithm Tips also consists of the actual email alerts sent to users, which are delivered as basic HTML. Here we strive for a display that is lightweight on information (DG1) with the goal of simply grabbing the user's attention and drawing them back to the web interface where they can see the details of the new leads found. The email alerts have the subject line "Algorithm Tips: New Leads Match Your Alert" with the body text in a template indicating the number of new leads and the alert filters: e.g. "Click here to see all 22 new leads matching your alert for (keyword filter: None; sources: Any)". When the user clicks the link they are taken to the corresponding filtered view of the database online. Because the state of the alert is stored in the URL parameters of the link, users can also share alerts (e.g. with collaborators) via other channels, or simply bookmark the alert in their browser to return to later (DG1).

Finally, the flags interface is very similar to the main database interface in that it displays a listing of leads that can be searched and filtered in the same ways (DG1). From there the user can also unflag a lead if they no longer want to keep a reference to it.

3.3.2 Information Design. Here we describe the visual information design and cues in a lead and connect that to the design goals articulated above. As can be seen in Figure 2 the lead is shown with a dominant title and summary of what the algorithm is and does, helping the user to assess relevance to their interests (DG1). A button on the title bar allows users to flag the lead, which highlights the item and also stores the item in the "flagged" set, making it easier to return later (DG1). A link to the original document where the algorithm was found is shown as well as a link to a cached copy of the document in case the original is no longer available online where it was initially found, which supports immediate follow-on research and initial verification activities (DG2). The text of the lead explains what keyword or key phrase was used to find the document, including a link to the actual query in case the user wants to see the original context of the search results (DG2). The date for when the original document was found and for when the lead was made available on the Algorithms Tips website is shown in order to provide context and also signal how recently the lead was found (DG2, DG3).

The "Additional Info" section of the lead highlights several pieces of metadata that are meant to support DG1 and DG3. In particular, the Jurisdiction, Source, and Main Topics fields are meant to help the journalist understand whether the lead is relevant to their interests or proximal to their location (DG1). The People and Organizations field is meant to signal whether there are key entities (i.e. "the power elite")that may be particularly newsworthy when judged by the journalist (DG3). For instance, if a document frequently mentions a powerful company or politician it may enhance the news value of reporting on the lead [45].

In the "Crowd Ratings" section the lead incorporates average ratings from the crowd for the four dimensions of newsworthiness that were crowdsourced (DG3). This includes a glanceable bar chart display so that as journalists are browsing or scrolling through leads, any outliers may more

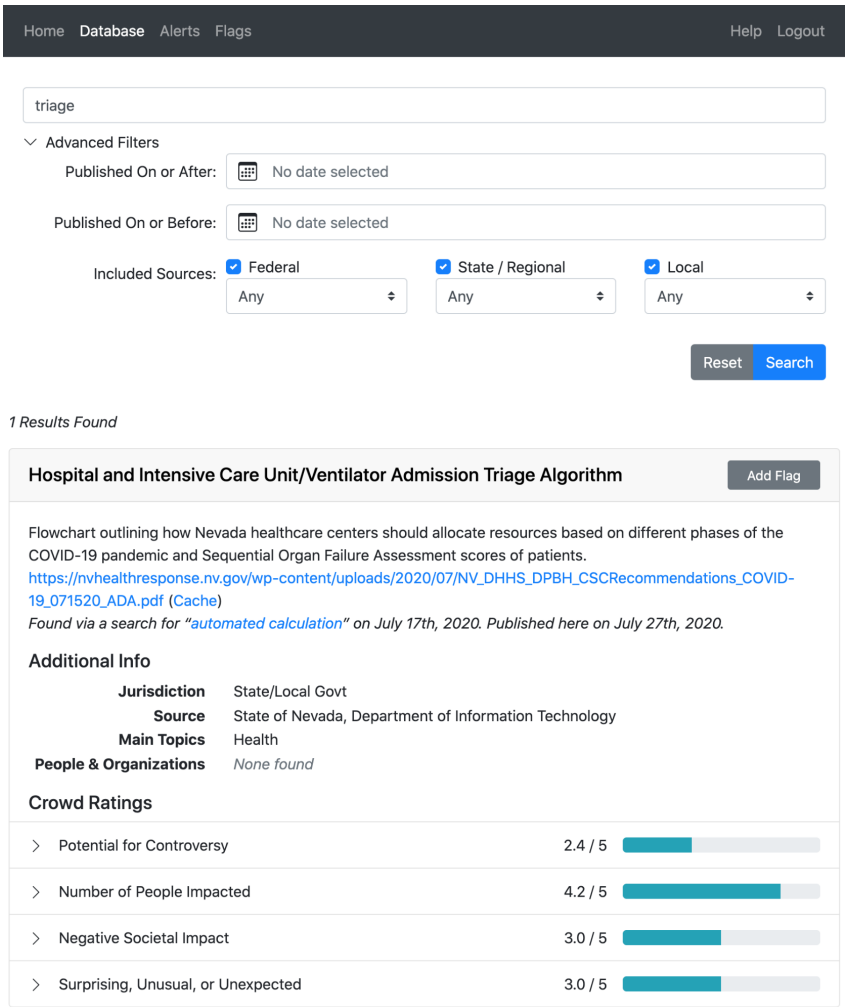


Fig. 2. The Algorithm Tips user interface showing search and filtering options as well as a single lead matching those options.

easily catch their attention (DG1). Clicking on any of the four factors will expand it to show a set of explanations and ratings from the individual crowd workers (DG2).

4 SYSTEM DEPLOYMENT AND STUDY

We conducted a field study to evaluate Algorithm Tips and its potential to support journalists in their identification and evaluation of potential news leads. Additionally we wanted to develop insights into how journalistic practices might be enabled and transformed by such news discovery tools more broadly. In our study we were motivated by the following questions: Can a tool like Algorithm Tips produce leads that are widely considered to be interesting to journalists? How can crowdsourcing be used to assist journalists in evaluating potential leads? How does such a tool fit the work routines of journalists and shape practices? We recruited participants to use Algorithm

Tips for an extended deployment period, after which we interviewed them to help study these questions.

4.1 Participants

Potential participants were recruited to the study using purposive sampling, defined as the "deliberate seeking out of participants with particular characteristics" [59]. This involved direct emails to journalists who had previously demonstrated interest in Algorithm Tips via a contact form that was posted on the system's home page, via posts on Twitter that solicited participation, and via in person solicitations at trade conferences. We sent email invitations to all the potential participants, with a link to a screening survey to determine eligibility. Eligible participants included professional journalists that worked in digital publications, newspapers, magazines, radio or TV, and had an interest in acquiring information about government algorithms. Additionally, since the documents that are collected and processed by Algorithm Tips are in English and refer to federal and local governments in the United States, we required participants to be English speakers and based in the U.S.

This process resulted in a total of eight participants who were currently employed in journalism and had a stated interest in finding out more about government algorithms due to their professional activities. The age of the participants ranges from 26 to 45. Six identified themselves as male and two as female. Four identified themselves as white, two as mixed-race or multiracial, one as Asian American, and one chose not to disclose their race. They work in a variety of positions and organizations (See Table 2), but tended to be oriented towards investigative or enterprise journalism which is characterized by extended time frames for reporting as well as longform presentations of content [42]. The expertise of our participants in covering algorithms in society adds to the context sensitivity, relevance, and external validity of our findings for actual expert practice. We compensated participants with a $25 gift card at the end of the study.

Table 2. Participant roles and organizations

Participant	Role	Organization
P1	Investigative data journalist	Digital-native news organization
P2	Data journalist	International news agency
P3	Data journalist	Asked not to share
P4	News application developer	Nonprofit investigative journalism organization
P5	Data journalist	Freelancer
P6	Data journalist	Local newspaper
P7	Senior reporter	Non-profit niche publication
P8	Investigative reporter	International news agency

4.2 Procedure

Participants were onboarded to the study in a 15-minute remote session, which began with us explaining the process of the study and asking them if they consented to participate. If they consented, we showed them the web application and explained how to use it. In this session, we also requested participants to set up at least one email alert that would let them know when the database was updated. Participants were then asked to use the web application at their own discretion for a period of eight weeks (from June 29 to August 22). In that period, they were welcome to search the database at any time and would receive any of the email alerts they had set up.

The workflow (described in Section 3.2) was configured to update the database on a weekly basis. To manage scale and the time budget afforded by available resources (~10 hrs / week) the internal expert, a third-year undergraduate journalism student, evaluated 40 documents per week (about 5% of the total); 35 were sampled based on the ranking of the relevance score from different jurisdictions (25 Federal; 5 State/Regional; 5 Local) and 5 were randomly sampled. Her work was also overseen and evaluated by a graduate student with 10 years of previous professional experience as a journalist, contributing an additional ~2 hrs / week of effort. The system was configured to scrape documents over a weekend, then the internal evaluator would spend the following week evaluating and annotating the sample of 40 documents, and leads were then uploaded to the database and sent to external journalists by the following Tuesday, resulting in a roughly 10 day lag from when documents were first discovered to when the enriched leads would be sent via email alerts.

At the end of the deployment period interviews were scheduled and conducted with each participant. In each semi-structured interview session we asked a series of questions that addressed both the usability of the system and its relevance to the journalists' workflow. We first asked the participants to describe their process in accessing and using Algorithm Tips. We followed that up with specific questions about the interface and the design of the tool, and whether the interface was effective for their needs. We asked them if there were any features that they think would be helpful to include in the tool. The next questions focused on the information retrieved by the tool and its relevance. We asked if the leads sent by Algorithm Tips were newsworthy. We asked whether the crowd ratings and the other metadata included with the leads (topics, source) had an impact on their evaluations. Next, we asked about their everyday usage of the tool: how often did they access Algorithm Tips, how much time did they spend looking at the leads, how did the leads fit into their broader workflow. Finally, we asked more reflective questions, such as their opinion as to whether Algorithm Tips or services like it could be adopted in news production, and whether they thought that such services introduce any ethical decisions into journalism practice.

All interviews were audio recorded with consent. The median interview lasted 38 minutes (Mean = 34 minutes). We acknowledge that a couple interviews were somewhat short as some respondents were under time pressure (as is common for journalists). Nonetheless, these were concise, definitive, and information-dense responses, and several interesting and rich comments are included in our findings from shorter interviews.

4.3 Analysis Method

Interviews were fully transcribed and then analyzed using an iterative qualitative thematic approach that was predominantly data-driven (i.e. themes emerging from the transcripts) though sensitized based on concepts that were used in framing and focusing questions [13, 38]. We began by open coding passages of the interviews, reading them closely to identify relevant thematic categories, and practicing constant comparison to identify similarities and contrasts as codes and themes were identified and triangulated across different respondents [39, 82]. Then, we developed memos to define these themes more precisely. Next, we connected related themes and identified axial codes that congealed around themes from the open coding, followed by additional memos for the axial codes. Throughout the process we practice analyst triangulation [66] between the co-authors by discussing open and axial codes and memos as they emerged and were refined and elaborated into themes. In the next section, we describe the themes that emerged via this analysis process.

5 FINDINGS

Based on the thematic analysis of the interviews three broad areas relevant to participants' use of computational news discovery tools emerged: lead interestingness, reporting practices, and

managing effort and attention. We detail these themes in the following subsections and discuss them more broadly in Section 6.

5.1 Lead Interestingness

The notion of "lead interestingness" captures how journalists worked to assess which entries in Algorithm Tips were interesting to them to pursue, and how the tool assisted them in this evaluative task. These reflections included considerations about what elements would make a lead intrinsically or contextually interesting, and how the current capabilities of Algorithm Tips impacted that evaluation.

5.1.1 General and Contextual Newsworthiness. Participants often reflected an understanding that some leads presented by Algorithm Tips were more interesting than others. For the journalists in our study, this interestingness was defined as newsworthiness; that is, the potential for a lead to become a news story that warrants publication. Participants related their conceptual interpretations of what newsworthiness meant both in general and to their own specific interests or the interests and scope of their organization (e.g. local news has a specific interest in the proximal impacts of algorithms). The contextual quality of newsworthiness, operating at several levels of influence (i.e. general, organizational, individual) complicates the clear delineation of leads as either newsworthy or not newsworthy:

> "There are kind of two components, I think, of judging something as newsworthy. One of them is the broad question of, you know, do you as an individual or you as a professional, independently of where you work, think that this is something that people should know about? And that's the first question. Can you answer that first in your head? And then there's the second component of it, which is, do you think that this is newsworthy for the outlet that you work at?" (P6)

Participants also noted that some leads were more clearly newsworthy in general, due to their reflection of specific qualities, such as the potential for controversy. In that respect, leads that mentioned algorithms that could cause bias or discrimination with a clear impact for individuals tended to receive more positive interest from participants. Commenting on a lead on Algorithm Tips that described a pretrial risk assessment tool, one of the participants pointed out that the algorithm was interesting because it was "about evaluating individuals, whereas ... a lot of the things that were on Algorithm Tips were like water quality, are not associated with individual rankings. Like, that's not where the controversy around algorithms comes from." (P7)

Participants also relayed the idea that some topics are more interesting than others, both in general for journalists and across beats, organizations, and even at different times. This interest in specific topics is contextualized by the organization and specific beat they work on:

> "A lot of the hits were sort of COVID related and fairly ... They were newsworthy, but they weren't the kind of thing I would write about. I think that a lot of them actually would be kind of interesting. You know, some governmental agency saying here's how we're going to reopen. Here's how we're going to judge whether or not we can do X, Y or Z with COVID. That's not the kind of algorithm that I'm super interested in at the moment." (P5)

Whether or not the participant had a specific interest in a topic also informed their strategies for exploring the leads. Participants that had a specific topic of interest used the Algorithm Tips design features to filter and search leads based on those topics. Participants that did not have a specific interest spent more time exploring the dataset and evaluating the leads. In the former cases, some participants expressed disappointment if their searches did not have any matches.

The interestingness of a lead could also be related to contextual factors around how easily a story based on that lead could be pursued. One participant indicated that the availability of public records (e.g. via FOIA or other public records laws) could enhance their interest; a lead mentioning a system in a jurisdiction that was easier to investigate because of public records availability made the lead more interesting and appealing:

> "California is a great place to have a tool to look into because their public records are pretty good. And like, if that's really something that's going to be done, I would think that's a good one to know about early, because there's a lot of reporting that can be done on it." (P1)

While considering the contextual implications of newsworthiness, and recognizing that each organization or reporter that uses Algorithm Tips might be interested in different types or topics of algorithms, some participants suggested that the system might benefit from a process to tailor the lead curation to specific, personalized needs. One participant specifically requested a personalized feedback loop system in which Algorithm Tips might learn what kind of leads are not interesting for that specific journalist: "So there is an add flag, right? Is there a way to tell the system whether to not show me these types of leads any more?" (P3). The idea was that this would help the journalist focus only on leads that were interesting to them in their particular context of reporting.

5.1.2 Definition Of "Algorithms". Relating to the distinct scope of the algorithmic accountability beat, practitioners also had different expectations in terms of what types of algorithms should be investigated. Some participants relayed a more specific understanding that "algorithms" in this context should mean software or even artificial intelligence. Some participants had issues with how "algorithms" were defined by Algorithm Tips (i.e. as not necessarily implemented computationally, see Section 3.2.3). This in turn made them consider the leads that pointed to non-software algorithms as not interesting:

> "I think my biggest concern with the tips that were presented was that they seem to reflect a definition of algorithm that was too broad. Certainly technically accurate, but too broad for a reporting perspective" (P2)

5.1.3 Usefulness Of Crowd Ratings. While participants were generally sympathetic to the idea of the crowdsourced newsworthiness ratings, they reported varying degrees of usefulness of those ratings and the associated written rationale. Different interpretations of the limitations of the crowd ratings led to an assortment of different strategies of use of those crowd ratings. Some would use the ratings as a filter by themselves, deciding to read closely all leads that had a particular minimum score: "As a ballpark, I just looked at it if the potential for controversy was more than four. And if it was more than four, I would look at what's going on in the dropped down comments" (P3). Others used them as checks for their own instincts, first evaluating the leads themselves and then checking if the raters agreed with them: "if I read ... a description and didn't think it was that interesting, and then saw scores that looked very high, I would be like, "oh, wait a second, maybe I need to go back and take a look at this thing" (P1). But in some cases these checks led to a sense of tension with their own judgement: "I would definitely look at them [but] I'm not sure how often I find myself in agreement with the ratings." (P4)

While most participants looked at the numerical ratings, most also reported that the written explanations were not so useful or insightful, either because they were redundant: "I didn't find the comments to be that insightful because they are often kind of conclusory what the ratings themselves seem" (P5) or because the explanations were too superficial:

> "Initially, the first time that I was doing this, I relied heavily on the ratings. Like, I would open up the ratings just to see, like, what the explanations were. And then after

> the first time, I realized that they didn't really provide me value ... It was clear that they weren't actually, like, really thinking deeply about this stuff." (P7)

Participants were generally aware that the crowd ratings were subjective interpretations and some appeared dismissive of them because of that. One exception was the rating of "Number of People Impacted," which some participants were inclined to give more importance as something implying a level of quantification. Ultimately, several participants affirmed that they thought that they should be the final judge, either for specific dimensions: "we can make a judgement call later on, whether it's surprising or unexpected" (P3), or more generally: "I did rely more on my own judgment to what's newsworthy as opposed to sort of the judgment of the crowd." (P6)

5.2 Reporting Practices

Journalists also shared a number of reflections on the role they expected to play when conducting algorithmic accountability reporting on leads generated by Algorithm Tips. We elaborate three primary areas that emerged relating to lead development, background knowledge, and ethical considerations.

5.2.1 Lead Development. Lead Development is the main process through which journalists turn leads into publishable news stories. Participants recognized the role of Algorithm Tips as a jumping off point, but also wanted the system to facilitate the next steps in the reporting process.

Jumping Off Point. Participants were cognizant and forthright about the idea that the leads provided by Algorithms Tips were a starting point, and that the central role of turning those leads into stories belongs to the journalists themselves. Participants overwhelmingly saw Algorithm Tips as the jumping off point to comprehend the algorithmic tools that are the targets of investigation, instead of a service that provides something close to a complete news story. P7 likened it to a feed of information: "I think of Algorithm Tips very much like Twitter in that it's just like a stream of potential stories," while P4 placed it in the broader context of tip services:

> "I think [Algorithm Tips] helps solve two problems, one of which is like the Rumsfeldian 'unknown unknowns.' You know, I don't know what I don't know. But I think it also can help address the sort of [situation that] I have some idea that something is going on, but I really don't know sort of how to start looking or where, who to talk to or what to read about this stuff ... in general, I think most reporters sort of value tip services of some kind." (P4)

Facilitating Next Steps. To make sure the investigation actually begins, reporters need information to determine if a lead is worth pursuing. While participants expect that the leads were the starting point, they also assume that the tool would provide some facilitation towards more information about the systems that are being mentioned. Some of them specifically highlighted the presence of a direct link to a government source document as a very positive feature that would help start their reporting: "It is nice that the link is directly there and you can just go right to where you got it from, like the source page and read about it, because that was definitely helpful." (P1)

Related to the usefulness of the presence of an official document is the need for specific information to evaluate how to proceed with the lead development. One participant questioned whether it would be possible to have more technical information about the inner workings of the systems that Algorithm Tips finds. This would help journalists start off their investigations by providing a better view of what the potential issues with the algorithms may be. Other participants mentioned how it would be useful to have more information about the source document, including the timeframe and context in which it was published. In general these bits of information would help journalists as they plan how (and whether) to proceed with their investigations. For instance, knowing the actual publication date, rather than when a document was *found* by the system could inform a next step,

as could the provenance and newness of the document with respect to other potentially related documents:

> "In my reporter's mind, I'd want to know, well, that's great that it was found on July 17, but is this last week news? Was it in fact published last week or is this, you know, something that was just published? (P8)

> "I have no idea where this PDF is from. Like, is it like part of a series of documents around a particular research project or is it a new initiative or anything like that? Like having that context would be helpful." (P7)

Participants also imagined ways that the leads they find in Algorithm Tips could be organized to fit their ongoing work processes. One participant requested a feature to have some mechanism that allows them to export the leads they found:

> That would not only allow me to go back in with, you know, the new ones I have having done some reporting over the last two weeks that advances my understanding of the topic, go back and see which ones are more interesting now, given my new accumulated knowledge, but also so that I could write a sweeping nut graph. (P8)

5.2.2 Background Information. As part of journalistic practice, journalists will often conduct background research including interviews and conversations with stake-holders and insiders to understand the beats they are covering [69]. Therefore, journalists recognize the usefulness of reading through leads and documentation of algorithmic systems that will not necessarily become news stories by themselves but which provide useful background information.

Echoing the process of pursuing background information, participants shared several comments related to the idea that the information Algorithm Tips surfaces can be interesting, but may not necessarily warrant a news story all on its own. Specifically, several participants related that just going through the list of algorithms allowed them to have a better understanding of how wide the scope for algorithmic tools is in government: "I mean, just the scope it has is informative because I didn't realize how much algorithms are being used for in government." (P3) and "My concept of how governments are using or potentially using algorithms or algorithmic approaches was expanded by the results." (P4). The ability of Algorithm Tips to pick up leads related to federal, state, and local levels of U.S. government may have contributed: "it's helpful to know how other levels and units of government approach some of these same problems" (P2).

5.2.3 Ethical Considerations. Journalists also see it as their role to resolve any ethical decisions or issues arising from reporting, including reporting that was instigated by computational lead production. But since participants saw Algorithm Tips as only the first step in an algorithmic accountability investigation, they also related that they had little concern about the potential ethical implications in using it.

> "The decision is ultimately up to me as the user whether to report on it, what to report about it, and that sort of thing. You know, to me, this is a more specialized search engine, and I don't necessarily have any ethical issues with that." (P2)

The one potential issue that participants saw with Algorithm Tips was that the tool itself might reproduce or create a skewed distribution of the kinds of algorithms there are in government use, leading to some filtering bias effect.

> "Perhaps if there was, a heavy reliance on this tool, there might be questions around like, what are the biases of this tool? What kinds of websites is it actually scanning to look for this stuff? And like what terms it is using to pick up stuff and how might that

affect its comprehensiveness and then affect the types of stories that I would write." (P7)

5.3 Managing Effort and Attention

Journalists have limited time and capacity to pursue leads, both in terms of the general time and labor constraints present in newsrooms, and in terms of temporal fluctuations of their needs and abilities. This mean practitioners have to evaluate whether or not the lead they found is worthy of additional labor, and how that labor can be facilitated by a tool such as Algorithm Tips.

5.3.1 How The Leads Fit Current Work Capacity. Several participants related having difficulty initiating algorithmic accountability reporting projects over the period of the study because they were either participating in other projects or had other limitations. This speaks not only to the general demands of journalism as a practice, but to temporal fluctuations of bandwidth and willingness to pursue new stories. As P1 expressed it,

> "I think if I was in the process of reporting a story or thing, I would check back and figure out which ones I wanted to make requests about or do additional reporting on. But that's not like where I am right now in the reporting process. (P1)

5.3.2 Reducing Effort. One issue related to the effort in pursuing leads is the effort in exploring them on Algorithm Tips in the first place. Participants reflected on what elements increased that workload, such as the volume of leads provided, and the labor involved in reading through them. Most participants indicated that they were eager to have more leads come through. Some mentioned that they realized that, as a whole, Algorithm Tips provides a sizable number of leads already, but that there were a few dimensions that they were specifically interested in for filtering, including topic of interest and current work priorities. The appropriate lead volume was understood to be related to lead relevance and scope to current needs.

Another pressure point in the effort of exploring leads was that, while participants recognized the importance of reading the governmental document that mentioned the algorithm to be investigated, sometimes this task was cumbersome, since they are often long and hard to parse documents.

> "And then there were some where there were like I remember there was like a couple where I was like, oh, I would like to learn more. So I clicked into the original documentation, but then the documentation was so long that I was like, it would take me too long to actually evaluate whether or not this is newsworthy." (P7)

Participants conveyed different strategies for navigating the leads to be more effective and reduce labor. While some participants relied on the email alerts triggered by keywords, some saw searching as the preferable interaction modality. One participant (P4) suggested linking more of the metadata in the interface so that a click would initiate a search in the tool, allowing for browsing and pivoting more easily. Another (P5) was interested in a way to put leads that were not interesting at the bottom of the list (the opposite of the 'flagging' feature the system currently supports), which would help focus the journalist's attention more towards leads that were interesting. In addition to filtering by search terms, some requested specific filtering capabilities for dimensions of interest to reduce the time needed to explore the leads in the tool.

> "If you could provide a drop down for selecting the state that would be very helpful because ... Once I narrowed it down to say, last month or something, and I get all of the results here, I want to then choose something like a checkbox of state." (P3)

6 DISCUSSION

In our findings we elaborate three interconnected themes that not only describe elements of usage of Algorithm Tips, but also echo concerns and issues of the journalism profession. First, Algorithm Tips and its use raises questions around the definition of interestingness or, as it is expressed in journalism, newsworthiness (Section 5.1). Then, with the leads appropriately evaluated and selected for further investigation, participants exercised aspects of their practice needed to turn leads into stories (Section 5.2.1). The use of Algorithm Tips echoes journalism practices not only in the process of specific investigations, but also in acquiring background information about their beats (Section 5.2.2) and in navigating ethical issues (Section 5.2.3). Finally, participants explained the challenges involved with actually putting investigative plans into practice, whether because of limited time or limited capacity to pursue new stories (Section 5.3).

Our findings indicate that the design goals that shaped the tool were largely helpful in supporting journalists' practices, while also noting areas in which features could be improved or added to further support those goals. Specifically, participants reported that Algorithm Tips provides a significant number of leads, that use of the tool could fit around their work and priorities, and that features such as searching and filtering helped facilitate the tool's sensitivity to available human effort and attention (DG1-Attention). Interviewees also related that they saw Algorithm Tips as a jumping-off point to their investigations, and that the tool provided contextual information and facilitation to verify and assess the quality of leads (DG2-Verification). Finally, participants reflected on how newsworthiness is contextual, and that the information provided by Algorithm Tips, including various aspects of metadata included in the interface, could help assess whether the leads were newsworthy in their specific context (DG3-Newsworthiness). Our evaluation also clarified several opportunities for improving the Algorithm Tips design to better inform evaluative decisions, facilitate lead development, and manage users' time and effort. Participants suggested improvements related to the navigation and filtering of leads (e.g. flagging irrelevant as well as relevant leads, more fine-grained state-level filters, and the need for better summarization or navigation of long documents), conceptual definition of the targets of the leads (in this case, government algorithms), and enhancements in the contextual information about the leads (e.g., document temporal context as well as improved quality of crowdsourced rating justifications).

In addition to suggesting these various opportunities for design refinements, the following subsections further elaborate several key ideas that emerge from the Algorithm Tips deployment findings, including issues related to how to support contextual journalistic decision making, newsworthiness evaluations and how to support them in news discovery tools, and sustainability of such tools as a subsidy for important public interest journalism.

6.1 Contextual Variance

One overarching finding is the importance of reporting context in driving evaluative decisions. That is, each practitioner is in a specific situation that informs their decision-making process, and, therefore, how they are going to use the tool at hand. In that respect, we identify two relevant aspects of reporting: what the journalist cares about; and what they can or are willing to do. The former is connected to the concept of algorithmic newsworthiness, or the idea that computational news discovery systems can be designed to support the configurability of assessments of interestingness, and the latter connects to the idea that such systems should take into account the expected investment of effort by journalists [25].

The question then is how a computational news discovery tool such as Algorithm Tips can be designed to fit that variety of journalistic contexts. This speaks to a broader issue of whether a system should support a specific use or general uses, and how to find the appropriate balance.

One example of the challenge of fitting into a variety of contexts that came up in this study is the tension generated by the differing definitions of "algorithm" in the context of algorithmic accountability reporting. Some participants had issues with the scope of "algorithms" defined by Algorithm Tips, which includes non-computational calculations that input data and output decisions. However, participants reported that their expectations were inspired by the colloquial use of the term, which implies some sort of software, if not artificial intelligence. At the same time, this predilection may be a tacit expression of the news value of "magnitude" with the assumption being that computational systems have a greater potential to scale up and impact more people. This tension raises interesting questions: Should the definition of algorithms be restricted in Algorithm Tips to match the expectation of reporters? Or should that distinction be explained and expressed clearly in the lead metadata to allow for users to filter according to their needs? Or perhaps the solution is somewhere in between, with customization—that is, should each user configure what definitions the system applies? Answering these questions entails different approaches to designing and supporting journalistic decision making, from realignment in the design process, to increased transparency, to better interactivity and configurability, all of which may have different implications for usability and for the computational models needed to support the system. In this case, the most straightforward design solution may be to simply include both definitions (one being a subset of the other) and allow users to filter leads according to the narrower definition if desired.

As we alluded to in the findings, some participants suggested that personalization might be a possible solution to adapting to a variety of contexts in a way that maximally prioritizes individual interests, current work needs, and perhaps organizational pressures. The proposition would be to create an explicit personalization mechanism [87] through which the user could provide feedback to the system (e.g. positive or negative ratings on relevance) so it would learn their preferences and sort and filter leads accordingly. This might help to increase the relevance of leads and align leads to journalists' idiosyncratic needs, and is reminiscent of a recent idea for helping to match science journalists to press releases [81]. Yet personalized information environments come with a host of other issues, often explored in end-user applications of news distribution, related to how such processes are biased and may shape information exposure [11, 58]. Based on our findings, it's already clear that personalization could inhibit different strategies for reporting and monitoring relevant information, such as by limiting the utility of the tool for facilitating background research. It also introduces new ethical concerns related to the biases of the tool that may suggest the need to develop new practices related to how editors setup, configure, and monitor how news discovery tools may be impacting coverage [23, 25]. Additional technical and human-centered evaluation work would need to assess whether the contextually-bound work of reporting could be effectively supported through such an approach. However, this is an interesting area for future work since the application of personalized algorithms for surfacing potentially newsworthy items is largely uncharted territory in terms of implications for the earlier newsgathering stages of gatekeeping processes [95].

6.2 Newsworthiness Judgements

In general we found that journalists were able to use the information provided by Algorithm Tips to inform their judgements of newsworthiness. There were some aspects of newsworthiness (e.g. bias and discrimination) that journalists were particularly interested in with respect to algorithmic decision making, which aligns with the reckoning U.S. society is currently undergoing with respect to discriminatory algorithms [6]. Given that some related work already suggests the utility of domain-specific news values in CND tools, we believe integrating domain-specific ratings (either crowdsourced or automated) into Algorithm Tips could additionally enhance journalistic evaluation of newsworthiness. Besides bias and discrimination, other domain-specific news factors suggested

by the literature on public sector algorithms might be whether an ADM impacts an individual's due-process rights, involves or is mediated by a human decision-maker, leads to unwarranted withholding of government entitlements, or leverages sensitive or personal information [10, 23, 49, 61]

News values and newsworthiness are conceptual distinctions promulgated by professionalized journalists to help assess interestingness [65], but are not intrinsic factors of news items. We observed this in journalists' assessments of the crowdsourced evaluation of leads, including the associated ratings and comments. While participants were generally interested in this feature, they came away from the study with varying levels of interest and use for that information, partially because the interpretation of the crowdworkers did not provide sufficient elaboration, and partially because the users disagreed, at times, with the evaluations provided. These disagreements expose an interesting professional tension between accepting journalists' own news judgements as subjective, while normatively preferring to defer to objectivity and the external attributes of a story in determining newsworthiness [17]. Confronting the subjectivity that was apparent to them in the crowdsourced ratings may have made such a tension more salient. An interesting area of future work would be to do more in-depth studies of how and why journalists might or might not incorporate crowd assessments of newsworthiness into their own judgements, including how the design of information interfaces may relate to uptake such as by signaling credibility.

While there are well-established, and somewhat well accepted, techniques such as AB headline testing, which directly incorporate audience feedback into publishing decisions [41], the use of audience (or crowd) evaluations of newsworthiness for lead identification is still new. Our findings open up several rich avenues for exploration in future work. One possibility would be to demand more of crowds by asking and prompting for more elaborate explanations of ratings so that the information would be more compelling to journalists. Another possibility would simply be to have more crowdraters evaluate each lead and aggregate high-level explanations for those ratings. Data collection could also be expanded to include written speculations about potential impacts which could function as user-generated content that could be further analyzed as a signal of newsworthiness and potential societal impact [96]. Yet another opportunity might be to re-define the "crowd" itself as the journalist's audience. If journalists were able to pre-test story leads with their specific audience in a process not unlike how AB testing works now with headlines, this could help guide them to news stories with greater relevance to their current audience while helping to focus their attention and reduce effort. Finally, we might consider developing machine learned models to evaluate various definitions of newsworthiness based on the ratings and texts collected from crowds, though this would in turn raise the stakes for algorithmic transparency and introduce new questions related to how journalists might perceive and incorporate such predicted scores.

6.3 Sustaining the Subsidy

We also wish to reflect on the broader premise of computational news discovery which motivated the development of Algorithm Tips, namely that such tools provide an information subsidy to journalists thereby reducing effort [71] and ultimately facilitating more reporting in the public interest [42, 43]. While our evaluation suggests that the tool was beneficial in informing journalists' practices of lead development, backgrounding, and reducing the effort involved with monitoring the government's use of algorithms to create useful jumping off points, the downstream effort required for investigative journalism (e.g. document inspection, public records requests, and talking to involved sources) is still substantial [83]. Future longitudinal evaluations (such as through reconstructive interviews [70]) will need to be undertaken to assess to what extent the benefits of tools like Algorithm Tips are born out in terms of leading to fully reported news stories that resulted from initial discovery via the tool.

A related question is the cost and sustainability of providing such a subsidy: even as external journalists may be enabled to pursue stories about algorithms, new work must be undertaken to operate Algorithm Tips. As outlined in Section 3.2, the information process implemented by Algorithm Tips is relatively intricate, involving the coordination of automated components, internal experts, and crowdsourced rating. In particular the use of internal experts involves careful training as well as periodic attention, which could limit the sustainability of the system as a long-term intervention to support external journalists. Our ongoing deployment of the system (which is not fully scaled up) costs approximately $800 per month to operate, which is dominated by labor expenses. Yet this human effort is unavoidable, since AI and automation are not at the level needed to operate at an acceptable level of quality without oversight. This in turn leads to a need to design clever information workflows that take advantage of automation where possible but also blend that with human effort to ensure the output meets professional expectations of accuracy and quality even though that may limit the benefits of automation to scale and speed [23, 83].

In this work we opted to deploy the use of automation as a complement to the internal evaluator to increase the yield of leads (i.e. scale) given a fixed time-budget for that evaluator (Option 1). Alternatively, one could imagine fixing the yield and instead reducing effort and time for the internal evaluator, thus lowering the cost and increasing sustainability (Option 2). Yet another model would fix the yield and allow the internal evaluator to reinvest time in other congruent activities, such as producing auxiliary content (e.g. blog posts, tweets) about the leads (Option 3). These non-exhaustive alternatives already suggest the importance of considering the output goal (e.g. labor saving vs. scale) in sustainably deploying such a tool. While we opted to deploy Option 1 for our study, a longer-term and more sustainable deployment might favor Option 2 or Option 3. Other alternatives might consider tuning the system for speed or latency which could enable different types of journalists beyond investigative or enterprise reporters and create interesting new challenges for designing workflows to meet those different temporal demands.

6.4 Limitations and Opportunities

We believe we achieved an adequate sample of practitioners to evaluate our system as we were able to recruit eight data journalists for a moderately long term study. Yet, as is the case with the journalism industry itself and data journalism in particular, the demographic distribution of the data journalists interviewed is not representative of the United States population.[7] One concern is the under-representation of women and limited racial diversity of journalists among the participants, which could have interesting but understudied implications with respect to the evaluation of newsworthiness of leads. Future work should explore and contrast newsworthiness decisions as well as the broader potential of CND tools for different types of individuals, types of reporters (e.g. investigative vs. daily beat reporters) or between different organizational or institutional contexts (e.g. national vs. local, commercial vs. nonprofit) with varying levels of resources.

Algorithm Tips collects relevant documents about government algorithms using online searches targeted at government websites. The effectiveness of this approach assumes that the government is posting documents openly online, which is not always the case, as different legal regimes or requirements and political contexts may impact the availability of government information on the web. We expect that the transparency afforded by public records requests will continue to be vital in investigating government power [10, 20] and are therefore considering how to integrate other sources of data such as relevant public information requests (e.g. via MuckRock) or repositories of public legal proceedings to help augment available information and facilitate our participants' interests in taking the initial steps of reporting.

[7] https://www.pewresearch.org/fact-tank/2018/11/02/newsroom-employees-are-less-diverse-than-u-s-workers-overall

This study was geographically limited to the United States, and therefore future work is needed to investigate its extensiblilty to other countries, where news values may differ and the structure of administrative society may demand different approaches to monitoring and filtering documents. Again, the reliance on freely published documents online may limit the applicability of the approach around the world where government behavior may vary in terms of openness. We are currently exploring internationalization of the tool, which would allow for a better understanding of different perspectives on algorithmic accountability reporting and the generalizability of such a tool across jurisdictions and cultures.

More generally we hope to explore whether a framework for the information process and workflow like the one that was developed for Algorithm Tips (See Figure 1) can be adapted for information gathering on subjects beyond government algorithms, such as health, science, or local news where information could be freely monitored online.

7 CONCLUSION

This work first presents the design of a prototype news discovery tool, Algorithm Tips, which was developed to help journalists find newsworthy leads related to the use of algorithms in the U.S. government. Drawing on prior research on computational news discovery tools we articulate the design goals and development of the information process and user interface of Algorithm Tips combining automated, expert, and crowdsourced ratings of news values to enrich documents and enable journalistic decision-making. We then presented an evaluation of the tool with eight professional journalists in an extended deployment in order to better understand how journalistic practices of news discovery are enabled and transformed by such a tool. Our findings suggest opportunities to improve computational news discovery tool designs and support those practices by better informing evaluative decisions, facilitating lead development, and managing user time and effort. Our findings furthermore expose several interesting new avenues for inquiry related to the incorporation of automated and crowdsourced ratings into news gatekeeping processes.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation via award IIS-1845460, and the Northwestern University Undergraduate Research Assistant Program (URAP). We thank Johnathan Smith for building the front-end of the system. This work wouldn't have been possible without the journalists who volunteered their time to use our system and participate in the research, thank you!

REFERENCES

[1] Elena Agapie, Jaime Teevan, and Andrés Monroy-Hernández. 2015. Crowdsourcing in the Field: A Case Study Using Local Crowds for Event Reporting. *Third AAAI Conference on Human Computation and Crowdsourcing* (Sept. 2015).

[2] Tanja Aitamurto. 2015. Crowdsourcing as a Knowledge-Search Method in Digital Journalism. *Digital Journalism* 4, 2 (May 2015), 280–297.

[3] Tanja Aitamurto, Mike Ananny, Chris W. Anderson, Larry Birnbaum, Nicholas Diakopoulos, Matilda Hanson, Jessica Hullman, and Nick Ritchie. 2019. HCI for Accurate, Impartial and Transparent Journalism: Challenges and Solutions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, 1–8.

[4] Omar Alonso. 2019. *The Practice of Crowdsourcing*. Morgan & Claypool.

[5] Franziska Badenschier and Holger Wormer. 2011. Issue Selection in Science Journalism: Towards a Special Theory of News Values for Science News? In *The Sciences' Media Connection –Public Communication and its Repercussions*. Springer Netherlands, Dordrecht, 59–85.

[6] Ruha Benjamin. 2019. *ARace After Technology: Abolitionist Tools for the New Jim Code*. Polity.

[7] Kenneth Benoit, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review* 110, 2 (May

2016), 278–295.
[8] Frank Bentley, Katie Quehl, Jordan Wirfs-Brock, and Melissa Bica. 2019. Understanding Online News Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, New York, USA, 1–11.
[9] Esha Bhandari and Rachel Goodman. 2017. Data Journalism and the Computer Fraud and Abuse Act: Tips for Moving Forward in an Uncertain Landscape *(Computation + Journalism Symposium)*.
[10] Hannah Bloch-Wehba. 2020. Access to Algorithms. *Fordham Law Review* 88 (2020).
[11] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3 (June 2013), 209–227.
[12] Petter Bae Brandtzaeg, Marika Lüders, Jochen Spangenberg, Linda Rath-Wiggins, and Asbjørn Følstad. 2015. Emerging Journalistic Verification Practices Concerning Social Media. *Journalism Practice* 10, 3 (2015), 323–342. https://doi.org/10.1080/17512786.2015.1020331
[13] Virginia Braun and Victoria Clarke. 2008. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (July 2008), 77–101.
[14] Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. 2014. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *In IEEE Transactions on Visualization and Computer Graphics (TVCG)* 20, 12 (2014), 2271–2280.
[15] Meredith Broussard. 2015. Artificial Intelligence for Investigative Reporting. *Digital Journalism* 3, 6 (Nov. 2015), 814–831.
[16] Madalina Busuioc. 2020. Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review* (2020). https://doi.org/10.1111/puar.13293
[17] Matt Carlson. 2017. Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. *New Media & Society* 8, 4 (May 2017), 146144481770668.
[18] Cary Coglianese and Lavi Ben Dor. 2020. AI in Adjudication and Administration: A Status Report on Governmental Use of Algorithmic Tools in the United States. *Brooklyn Law Review* (2020).
[19] Sarah Cohen, James T. Hamilton, and Fred Turner. 2011. Computational journalism. *Commun. ACM* 54, 10 (Oct. 2011), 66–71.
[20] David Cuillier and Charles N. Davis. 2011. *The Art of Access: Strategies for Acquiring Public Records.* CQ Press, Washington DC.
[21] Dharma Dailey and Kate Starbird. 2014. Journalists as Crowdsourcerers: Responding to Crisis by Reporting with a Crowd. *Computer Supported Cooperative Work (CSCW)* 23, 4 (Dec. 2014), 445–481.
[22] Nicholas Diakopoulos. 2015. Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital Journalism* 3, 3 (2015), 398–415.
[23] Nicholas Diakopoulos. 2019. *Automating the news: How algorithms are rewriting the media.* Harvard University Press.
[24] Nicholas Diakopoulos. 2019. Towards a Design Orientation on Algorithms and Automation in News Production. *Digital Journalism* 7, 8 (Nov. 2019), 1180–1184.
[25] Nicholas Diakopoulos. 2020. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. *Digital Journalism* (2020), 1–23.
[26] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the CHI conference on human factors in computing systems*. ACM, 2451–2460.
[27] Nicholas Diakopoulos, Madison Dong, Leonard Bronner, and Jeremy Bowers. 2020. Generating Location-Based News Leads for National Politics Reporting. In *Proc. Computation + Journalism Symposium*.
[28] Nicholas Diakopoulos, Sergio Goldenberg, and Irfan Essa. 2009. Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, 799.
[29] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*. 115–122.
[30] Fernando José Fradique Duarte, Óscar Mortágua Pereira, and Rui L. Aguiar. 2019. Framework for the Discovery of Newsworthy Events in Social Media. *International Journal of Organizational and Collective Intelligence (IJOCI)* 9, 3 (2019), 45–62. https://doi.org/10.4018/ijoci.2019070103
[31] David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies.* Technical Report.
[32] Nel Escher and Nikola Banovic. 2020. Exposing Error in Poverty Management Technology: A Method for Auditing Government Benefits Screening Tools. *PACM Human-Computer Interaction* 4, CSCW1 (2020). https://doi.org/10.1145/3392874

[33] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

[34] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. 2018. Falling for Fake News: Investigating the Consumption of News via Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173950

[35] Luciano Floridi. 2020. Artificial Intelligence as a Public Service: Learning from Amsterdam and Helsinki. *Philosophy & Technology* 33, 4 (10 2020), 541–546. https://doi.org/10.1007/s13347-020-00434-3

[36] Asbjørn William Ammitzbøll Flügge, Thomas Hildebrandt, and Naja Holten Møller. 2020. Algorithmic Decision Making in Public Services: A CSCW-Perspective. In *GROUP '20: Companion of the 2020 ACM International Conference on Supporting Group Work*. 111–114. https://doi.org/10.1145/3323994.3369886

[37] Floyd J. Fowler. 1995. *Improving survey questions : design and evaluation.* Sage Publications.

[38] Graham R Gibbs. 2007. *Analyzing qualitative data.* Sage London, England.

[39] Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research.* Routledge.

[40] Lucas Graves. 2018. *Understanding the Promise and Limits of Automated Fact-Checking.* Technical Report. Reuters Institute for the Study of Journalism.

[41] Nick Hagar and Nicholas Diakopoulos. 2019. Optimizing Content with A/B Headline Testing: Changing Newsroom Practices. *Media and Communication* 7, 1 (2019), 117–127. https://doi.org/10.17645/mac.v7i1.1801

[42] James T. Hamilton. 2016. *Democracy's Detectives.* Harvard University Press.

[43] James T. Hamilton and Fred Turner. 2009. *Accountability through algorithm: Developing the field of computational journalism.* Center for Advanced Study in the Behavioral Sciences.

[44] Tony Harcup and Deirdre O'Neill. 2001. What Is News? Galtung and Ruge revisited. *Journalism Studies* 2, 2 (May 2001), 261–280.

[45] Tony Harcup and Deirdre O'Neill. 2016. What is news? News values revisited (again). *Journalism Studies* 23, 1 (March 2016), 1–19.

[46] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *Proc. Knowledge Discovery and Datamining KDD*.

[47] Melinda McClure Haughey, Meena Devii Muralikumar, Cameron Wood, and Kate Starbird. 2020. On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (2020).

[48] Melanie Kellar, Carolyn Watters, and Kori M Inkpen. 2007. An exploration of web-based monitoring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM Press, New York, New York, USA, 377–386.

[49] Alayna Kennedy, Daphne Coates, and Katelyn Lindquist. 2020. Auditing Government AI: How to assess ethical vulnerability in machine learning. In *Workshop on Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

[50] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: Crowdsourcing Complex Work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/2047196.2047202

[51] Tomoko Komatsu, Marisela Gutierrez Lopez, Stephann Makri, Colin Porlezza, Glenda Cooper, Andrew MacFarlane, and Sondess Missaoui. 2020. AI should embody our values: Investigating journalistic values to inform AI technology design. In *Proceedings of the Nordic Conference on Human-Computer Interaction*. https://doi.org/10.1145/3419249.3420105

[52] Raymond Liaw, Ari Zilnik, Mark Baldwin, and Stephanie Butler. 2013. Maater: Crowdsourcing to Improve Online Journalism. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13*. 2549–2554. https://doi.org/10.1145/2468356.2468828

[53] Fabienne Lind, Maria Gruber, and Hajo G. Boomgaarden. 2017. Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs. *Communication Methods and Measures* 11, 3 (May 2017), 191–209.

[54] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Sameena Shah, Robert Martin, and John Duprey. 2017. Reuters tracer: Toward automated news production using large scale social media data. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1483–1493.

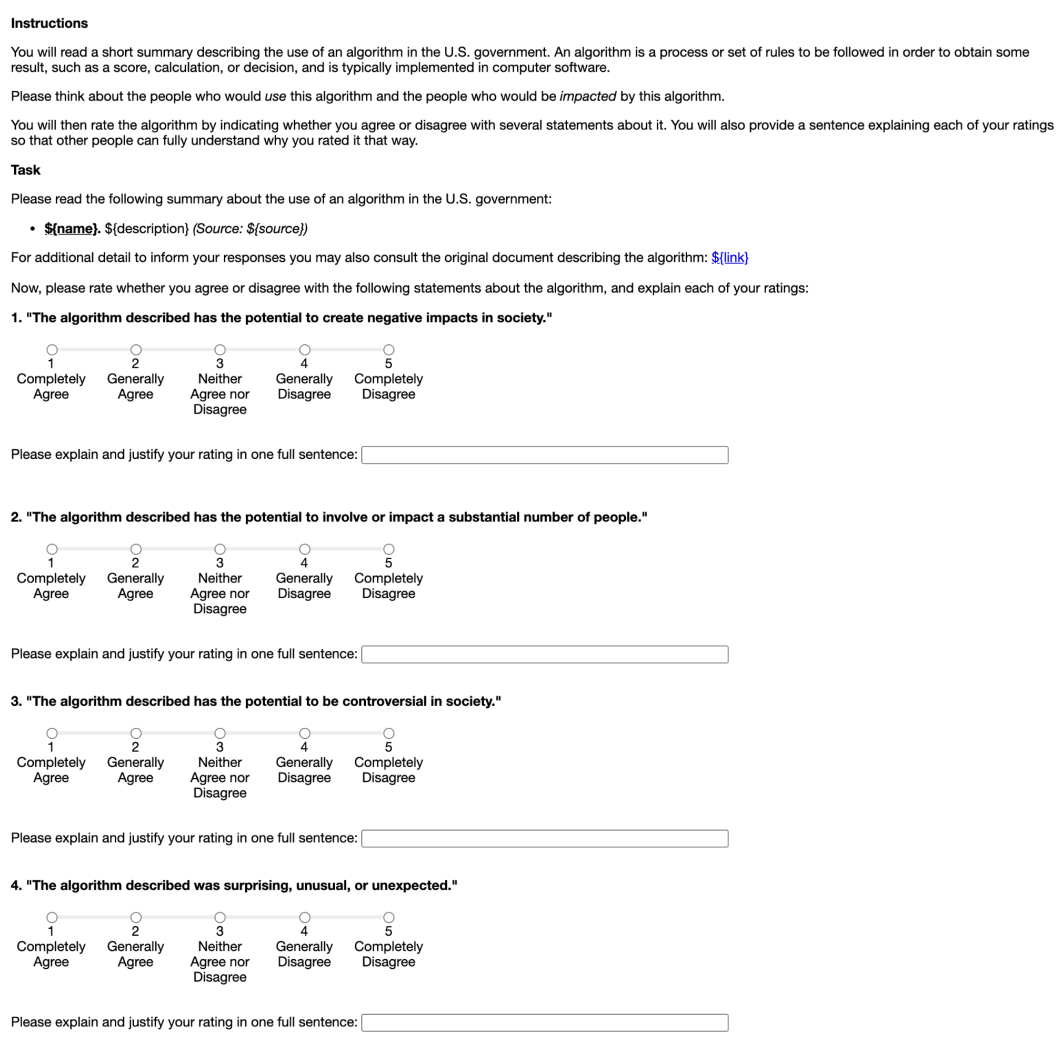
[55] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. 2016. Finding the news lead in the data haystack: Automated local data journalism using crime data. In *Computation + Journalism Symposium*.

[56] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the News: Digital Creativity Support for Journalists. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–11.

[57] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. *Fourth AAAI Conference on Human Computation and Crowdsourcing* (Sept. 2016).
[58] Cristina Monzer, Judith Moeller, Natali Helberger, and Sarah Eskens. 2020. User Perspectives on the News Personalisation Process: Agency, Trust and Utility as Building Blocks. *Digital Journalism* (2020), 1–21.
[59] Janice Morse. 2004. Purposive Sampling. In *The SAGE Encyclopedia of Social Science Research Methods*, Michael Lewis-Beck, Alan Bryman, and Tim Futing Liao (Eds.). Sage Publications, Inc., Thousand Oaks ,CA.
[60] Enrico Motta, Enrico Daga, Andreas L. Opdahl, and Bjørnar Tessem. 2020. Analysis and Design of Computational News Angles. *IEEE Access* 8 (2020).
[61] Naja Holten Møller, Irina Shklovski, and Thomas T Hildebrandt. 2020. Shifting concepts of value: Designing algorithmic decision-support systems for public services. In *Proceedings of the Nordic Conference on Human-Computer Interaction*. 1–12. https://doi.org/10.1145/3419249.3420149
[62] Armineh Nourbakhsh, Quanzhi Li, Xiaomo Liu, and Sameena Shah. 2017. "Breaking" Disasters - Predicting and Characterizing the Global News Value of Natural and Man-made Disasters.. In *Data Science + Journalism Workshop*.
[63] Changhoon Oh, Jinhan Choi, Sungwoo Lee, SoHyun Park, Daeryong Kim, Jungwoo Song, Dongwhan Kim, Joonhwan Lee, and Bongwon Suh. 2020. Understanding User Perception of Automated News Generation System. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13.
[64] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1114–1125.
[65] Perry Parks. 2018. Textbook News Values: Stable Concepts, Changing Choices. *Journalism & Mass Communication Quarterly* 96, 3 (Oct. 2018), 784–810.
[66] Michael Quinn Patton. 2014. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* (4 ed.). Sage Publications.
[67] William Perrin. 2017. Local News Engine: Can the machine help spot diamonds in the dust? In *Data Journalism Past, Present, Future*, John Mair, Richard Lance Keeble, Megan Lucero, and Martin Moore (Eds.). Abramis academic publishing.
[68] Alicja Piotrkowicz, Vania Dimitrova, and Katja Markert. 2017. Automatic Extraction of News Values from Headline Text. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, 64–74.
[69] David Randall. 2011. *The Universal Journalist* (4 ed.). PlutoPress.
[70] Zvi Reich. 2006. The Process Model of News Initiative. *Journalism Studies* 7, 4 (2006), 497–514. https://doi.org/10.1080/14616700600757928
[71] Zvi Reich and Yigal Godler. 2014. A Time of Uncertainty: The effects of reporters' time schedule on their work. *Journalism Studies* 15, 5 (2014), 607–618. https://doi.org/10.1080/1461670x.2014.882484
[72] Daniel Riffe, Stephen Lacy, and Frederick Fico. 2005. *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (2nd ed.). Lawrence Erlbaum, Mahwah, NJ, USA.
[73] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms used within the U.S. Child Welfare System. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3313831.3376229
[74] Michael Schudson. 2011. *The Sociology of News* (2 ed.). W. W. Norton and Company, Inc.
[75] Ida Schultz. 2007. The Journalistic Gut Feeling: Journalistic Doxa, News Habitus and Orthodox News Value. *Journalism Practice* 1, 2 (June 2007), 190–207.
[76] Raz Schwartz, Mor Naaman, and Rannie Teodoro. 2015. Editorial algorithms: Using social media to discover and report local news. In *International Conference on Web and Social Media*.
[77] M. Shearer, B. Simon, and C. Geiger. 2014. Datastringer: easy dataset monitoring for journalists. In *Proceedings Symposium on Computation + Journalism*.
[78] Pamela J. Shoemaker and Timothy Vos. 2009. *Gatekeeping Theory*. Routledge, New York, NY, USA.
[79] Katie A. Siek, Gillian R. Hayes, Mark W. Newman, and John C. Tang. 2014. Field Deployments: Knowing from Using in Context. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (Eds.). 267–289. https://doi.org/10.1007/978-1-4939-0378-8_6
[80] C. Estelle Smith, Eduardo Nevarez, and Haiyi Zhu. 2020. Disseminating Research News in HCI: Perceived Hazards, How-To's, and Opportunities for Innovation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 1–13.
[81] C. Estelle Smith, Xinyi Wang, Raghav Pavan Karumur, and Haiyi Zhu. 2018. [Un]Breaking News: Design Opportunities for Enhancing Collaboration in Scientific Media Production. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173955

[82] Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research.* Sage Publications.
[83] Jonathan Stray. 2019. Making Artificial Intelligence Work for Investigative Journalism. *Digital Journalism* 7, 8 (2019), 1–22. https://doi.org/10.1080/21670811.2019.1630289
[84] James Surowiecki. 2004. *The Wisdom of Crowds : Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Little Brown, London, England.
[85] Keren Tenenboim-Weinblatt and Motti Neiger. 2018. Temporal affordances in the news. *Journalism* 19, 1 (2018), 37–55. https://doi.org/10.1177/1464884916689152
[86] Neil Thurman. 2019. Computational Journalism. In *The Handbook of Journalism Studies* (2nd ed.), Karin Wahl-Jorgensen and Thomas Hanitzsch (Eds.).
[87] Neil Thurman and Steve Schifferes. 2012. The Future of Personalization at News Websites. *Journalism Studies* 13, 5-6 (10 2012), 775 – 790. https://doi.org/10.1080/1461670x.2012.664341
[88] Peter Tolmie, Rob Procter, David William Randall, Mark Rouncefield, Christian Burger, Geraldine Wong Sak Hoi, Arkaitz Zubiaga, and Maria Liakata. 2017. Supporting the use of user generated content in journalistic practice. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 3632–3644.
[89] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15.
[90] Daniel Trielli, Jennifer Stark, and Nicholas Diakopoulos. 2017. Algorithm Tips: A Resource for Algorithmic Accountability in Government. In *Proc. Computation + Journalism Symposium*.
[91] Damian Trilling, Petro Tolochko, and Björn Burscher. 2016. From Newsworthiness to Shareworthiness: How to Predict News Sharing Based on Article Characteristics. *Journalism & Mass Communication Quarterly* 94, 1 (March 2016), 38–60.
[92] Michael Veale and Irina Brass. 2019. Administration by Algorithm? Public Management meets Public Sector Machine Learning. In *Algorithmic Regulation*. https://doi.org/10.31235/osf.io/mwhnb
[93] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.31235/osf.io/8kvf4
[94] Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. 2019. GroundTruth: Augmenting Expert Image Geolocation with Crowdsourcing and Shared Representations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 1–30.
[95] Julian Wallace. 2018. Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. *Digital Journalism* 6, 3 (2018), 274–293.
[96] Yixue Wang and Nicholas Diakopoulos. 2021. Journalistic Source Discovery: Supporting The Identification of News Sources in User Generated Content. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15.
[97] Patricia Wei, Holden S. Foreman, Erin D. Bennett, and Christopher H. Stock. 2020. Agenda Watch: personalized email alerts for public meetings. In *Computation + Journalism Symposium*.
[98] Meng-Han Wu and Alexander James Quinn. 2017. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. *Fifth AAAI Conference on Human Computation and Crowdsourcing* (Sept. 2017).
[99] Barbie Zelizer. 2018. Epilogue: Timing the study of news temporality. *Journalism* 19, 1 (2018), 111–121. https://doi.org/10.1177/1464884916688964
[100] Arkaitz Zubiaga. 2019. Mining social media for newsgathering: A review. *Online Social Networks and Media* 13 (2019), 100049.
[101] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PloS one* 11, 3 (03 2016), e0150989. https://doi.org/10.1371/journal.pone.0150989

A APPENDIX



Instructions

You will read a short summary describing the use of an algorithm in the U.S. government. An algorithm is a process or set of rules to be followed in order to obtain some result, such as a score, calculation, or decision, and is typically implemented in computer software.

Please think about the people who would *use* this algorithm and the people who would be *impacted* by this algorithm.

You will then rate the algorithm by indicating whether you agree or disagree with several statements about it. You will also provide a sentence explaining each of your ratings so that other people can fully understand why you rated it that way.

Task

Please read the following summary about the use of an algorithm in the U.S. government:

- **${name}.** ${description} *(Source: ${source})*

For additional detail to inform your responses you may also consult the original document describing the algorithm: ${link}

Now, please rate whether you agree or disagree with the following statements about the algorithm, and explain each of your ratings:

1. "The algorithm described has the potential to create negative impacts in society."

1 Completely Agree | 2 Generally Agree | 3 Neither Agree nor Disagree | 4 Generally Disagree | 5 Completely Disagree

Please explain and justify your rating in one full sentence:

2. "The algorithm described has the potential to involve or impact a substantial number of people."

1 Completely Agree | 2 Generally Agree | 3 Neither Agree nor Disagree | 4 Generally Disagree | 5 Completely Disagree

Please explain and justify your rating in one full sentence:

3. "The algorithm described has the potential to be controversial in society."

1 Completely Agree | 2 Generally Agree | 3 Neither Agree nor Disagree | 4 Generally Disagree | 5 Completely Disagree

Please explain and justify your rating in one full sentence:

4. "The algorithm described was surprising, unusual, or unexpected."

1 Completely Agree | 2 Generally Agree | 3 Neither Agree nor Disagree | 4 Generally Disagree | 5 Completely Disagree

Please explain and justify your rating in one full sentence:

Fig. 3. Instructions and layout for the crowdsourced newsworthiness assessment task.

Received January 2021; revised April 2021; accepted July 2021