# Agency and Amplification: A Comparison of Manual and Computational Thematic Analyses by Public Health Researchers



ROBERT P. GAUTHIER, University of Waterloo, Canada
CATHERINE PELLETIER, Université Laval, Canada
LAURIE-ANN CARRIER, Université Laval, Canada
MAUDE DIONNE, Institut National de Santé Publique du Québec, Canada
ÈVE DUBÉ, Université Laval, Canada and Institut National de Santé Publique du Québec, Canada
SAMANTHA MEYER, University of Waterloo, Canada
JAMES R. WALLACE, University of Waterloo, Canada

Computational techniques offer a means to overcome the amplified complexity and resource-intensity of qualitative research on online communities. However, we lack an understanding of how these techniques are integrated by researchers in practice, and how to address concerns about researcher agency in the qualitative research process. To explore this gap, we deployed the Computational Thematic Analysis Toolkit to a team of public health researchers, and compared their analysis to a team working with traditional tools and methods. Each team independently conducted a thematic analysis of a corpus of comments from Canadian news sites to understand discourses around vaccine hesitancy. We then compared the analyses to investigate how computational techniques may have influenced their research process and outcomes. We found that the toolkit provided access to advanced computational techniques for researchers without programming expertise, facilitated their interaction and interpretation of the data, but also found that it influenced how they approached their thematic analysis.

CCS Concepts: • **Computing methodologies**; • **Human-centered computing** → **Empirical studies in HCI**; **Collaborative and social computing theory, concepts and paradigms**;

Additional Key Words and Phrases: case study, field deployment, comparison, thematic analysis, computational methods

**ACM Reference Format:**
Robert P. Gauthier, Catherine Pelletier, Laurie-Ann Carrier, Maude Dionne, Ève Dubé, Samantha Meyer, and James R. Wallace. 2023. Agency and Amplification: A Comparison of Manual and Computational Thematic Analyses by Public Health Researchers. *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 2 (January 2023), 22 pages. https://doi.org/10.1145/3567552

Authors' addresses: Robert P. Gauthier, rpgauthier@uwaterloo.ca, University of Waterloo, Waterloo, Ontario, Canada; Catherine Pelletier, catherine.pelletier@crchudequebec.ulaval.ca, Université Laval, Québec, Québec, Canada; Laurie-Ann Carrier, laurie-ann.carrier.1@ulaval.ca, Université Laval, Québec, Québec, Canada; Maude Dionne, maude.dionne@inspq.qc.ca, Institut National de Santé Publique du Québec, Québec, Québec, Canada; Ève Dubé, Eve.Dube@inspq.qc.ca, Université Laval, Québec, Québec, Canada and Institut National de Santé Publique du Québec, Québec, Québec, Canada; Samantha Meyer, samantha.meyer@uwaterloo.ca, University of Waterloo, Waterloo, Ontario, Canada; James R. Wallace, james.wallace@uwaterloo.ca, University of Waterloo, Waterloo, Ontario, Canada.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2573-0142/2023/1-ART2 $15.00
https://doi.org/10.1145/3567552

## 1 INTRODUCTION

Recent progress in artificial intelligence and machine learning research has prompted the human-computer interaction research community to question how computational techniques should be used to perform data science. They've noted, for instance, that computational techniques converge with qualitative research activities like coding and theme creation, and can help researchers overcome human limitations like the time required to (manually) study massive online data sets [6, 36, 43]. While the theoretical benefits of these partnerships are compelling, we have a limited understanding of whether they are ultimately beneficial, how they might improve research practices, or whether they may have some unintended side-effects or limitations.

To date, the human-computer interaction research community has developed prototype tools (e.g., [7, 22]), and spoken to qualitative researchers about their perceptions, beliefs, and attitudes towards AI (e.g., [20, 26] ). Critical computing researchers have also begun integrating qualitative traditions, such as reflexivity and context awareness, into quantitative computational methods to mitigate bias, respect user groups, and consider social conditions during the development of data sets and models [2, 16, 39, 44, 48]. However, we currently lack hands-on experience with these tools, and a sense of how they might be used in practice, particularly by researchers without programming experience.

To explore how these partnerships play out, we asked two teams of public health researchers to independently perform thematic analyses on the same data set; one performed 'manually' using a traditional process, and one performed 'computationally' using the Computational Thematic Analysis (CTA) Toolkit [22]. Each team started with the same set of 613,666 comments from English Canadian news sites and independently performed an inductive thematic analysis to answer research questions about the perception of COVID-19 during the period of January – July 2020. After both teams had finished, we compared each team's process and results.

In presenting the results of our case study we describe how each analysis followed a different inductive and iterative process — the manual analysis team used a bottom-up process defined by the human activities, whereas the computational analysis team interacted used a top-down process defined by the toolkit modules. Despite these differences, each analysis produced coding trees with similar, overlapping themes. Reflecting on these results, we discuss implications of qualitative researchers' use of the toolkit to perform thematic analysis:

(1) The toolkit lowered the threshold to use computational techniques, and enabled non-programmers to conduct their qualitative research;
(2) It simultaneously 'raised the ceiling' [29] of their efforts, and facilitated their interaction and interpretation of the large data set;
(3) However, we also show how use of the toolkit influenced their research process, and discuss implications of this influence for future research.

## 2 BACKGROUND

Our work is situated at the intersection of three active research areas: thematic analysis, computational social science, and toolkit research. In particular, we examine how techniques from computational social science (e.g., [8, 50]) can be used to address known challenges in applying qualitative research methods like reflexive thematic analysis to large, online data sets [13, 18]. To do so, we were particularly interested in understanding how toolkit-based research might help us understand some of the implications of deploying computational tools to domain experts in a practical setting (e.g., [22, 29]). We now outline related research that describes: (1) challenges in extending traditional qualitative methods to online data, (2) the potential benefits of computational

techniques for these analyses, and (3) how toolkit-based research can be used to explore their utility through real world deployments.

### 2.1 Thematic Analysis

Thematic analysis is a flexible research method that is used to develop, analyze, and report qualitative themes present within data [11, 13]. Being a flexible method means there is not one correct approach to performing a thematic analysis; instead there are multiple sets of adaptable guidelines aligned under three schools: reflexive, codebook, and coding reliability. Further, within these schools, thematic analyses can be performed with different levels of formality, ranging from adhering closely to an established procedure, such as *reflexive [13]* or *codebook [11]*, to less formal approaches like *pragmatic* thematic analysis [3].

A consequence of this flexibility is that planning and performing a thematic analyses is complex. Researchers need to consider multiple interconnected aspects of their research such as: its context and objectives; the researcher's position, assumptions, and experiences; research constraints, such as time, money, and people; and which reliability mechanisms are appropriate [14, 34]. These complexities are compounded by the iterative nature of thematic analysis which requires researchers to repeatedly read and code data, and develop and refine themes which contextually integrate their experience and practical knowledge [11, 13].

And when working with online communities, this complexity is even more severe. Researchers need to consider their data's origin and scale . They also need to navigate trade-offs between the amount of data included in their analysis and the resources required to analyze it [13, 18].

A common strategy to manage these challenges is sampling to reduce the amount of data processed for analysis. For instance, researchers frequently use random selection [4] or convenience sampling, such as a date-window [1, 18, 23, 42], to obtain a sample that is small enough for human analysis. However, these sampling techniques risk discarding interesting data before they can be considered by expert researchers [33]. Thus, researchers have turned to computational strategies, such as purposive sampling [25], to mitigate these challenges while simultaneously enabling them to engage with data in detail and interpret themes from large data sets.

### 2.2 Computational Social Science

The human-computer interaction research community has recently begun to explore 'convergences' between machine learning and qualitative methods: the collection of empirical evidence, iterative interaction with data, and the use of different lenses to interpret it [6, 17, 36, 43]. To date, these convergences have manifested through research that explores how computational techniques like topic modelling (e.g., [8]) and exploratory search [49] can support qualitative researchers in identifying interesting samples [21, 25] and latent topics [32, 40]. These techniques are particularly compelling because they are interpretive, and can be grounded in researchers' expertise and practical knowledge [33].

However, qualitative researchers also do not want computational techniques to simply automate their interaction with the data; they want to maintain autonomy, intimacy, and ownership of their analysis [26]. There are also open questions around how these influences manifest because computational techniques have been primarily developed and validated by computational researchers without sufficient input from qualitative researchers with domain expertise [5].

Moreover, the integration of computational social science techniques into qualitative research is incomplete [5, 26]. While the aforementioned research has developed narrow technical prototypes that explore individual computational techniques, they have yet to deeply explore their use in practice, by domain experts. Developed tools also frequently rely on programming knowledge

and assume a process grounded in data science, which makes it difficult to combine and validate technique use within a qualitative research pipeline [5].

To address this gap, the human-computer interaction research community has called for work that bridges the gaps between the computational and qualitative communities to establish common ground and explore where and how computational techniques can provide meaningful value to the social science [5, 17]. Our work contributes to closing these gaps by deploying a toolkit to an existing research group actively engaged in addressing a pressing real world public health research question.

### 2.3 Toolkit-based Research

In human-computer interaction research, toolkits provide opportunities to explore a 'bold vision of the future' and enable access to new solution spaces [29]. Toolkits can be developed and deployed by researchers to empower new audiences through access to tools; to explore concerns and research gaps identified in the literature, like those surrounding agency [5, 26]; and to understand how they integrate with current practice. During our toolkit deployment, we focus on two toolkit evaluation strategies identified by Ledo et al. [29]: *demonstrations* and *usage*.

Demonstrations explore what can be done with a toolkit and enable researchers to describe which paths of least resistance it facilitates [29]. They use methods like case studies in real world contexts to describe how toolkits can be used as unexpected situations occur (e.g., [31, 45]). They help to identify thresholds and ceilings — a person's ability to get started, and how much they can achieve with a toolkit [29] — both of which contribute to understanding how and where a toolkit can be used in complex solution spaces [37].

Usage evaluations explore who can use a toolkit and frequently include end users as valuable stakeholders [29]. They take advantage of methods like take-home studies to understand how stakeholders appropriate and use toolkits over time while developing new workflows (e.g., [9, 27]). Similarly, usage evaluations provide opportunities to reflect on how people's behaviour differs with the introduction of a toolkit (e.g., [10, 24]). They are useful for exploring complex design spaces, like thematic analysis of large data sets, where integration of computational techniques has been theorized but is not yet fully established.

In this work, we explore both demonstration and usage through a case study of holistic use of a toolkit that supports thematic analysis. By deploying the toolkit with a team of domain experts, we sought to better understand its thresholds and ceilings, and how it can be used to solve complex tasks. We also explored how the toolkit may influence their analysis process and outcomes.

## 3 THE COMPUTATIONAL THEMATIC ANALYSIS TOOLKIT

The Computational Thematic Analysis Toolkit [22] was designed to enable non-programmers to use computational techniques to perform thematic analysis of online community data. In developing their toolkit, Gauthier and Wallace created a cohesive, visual interface that integrates Braun and Clark's [13] reflexive thematic analysis phases with tasks common to data science (Figure 1). To support these various activities, the toolkit comprises interconnected modules that researchers can freely move between as they iteratively familiarize themselves with their data, interpret it, and reflect on their findings:

- The **Data Cleaning & Filtering module** enables researchers to visualize collected data and how NLP techniques may interpret it through fields like included and removed tokens, part of speech, and NLP summaries, such as frequency and TF-IDF range. This module also enables researchers to interactively review and change the filtering rules being applied to the data, and the impact of those changes on the above fields.

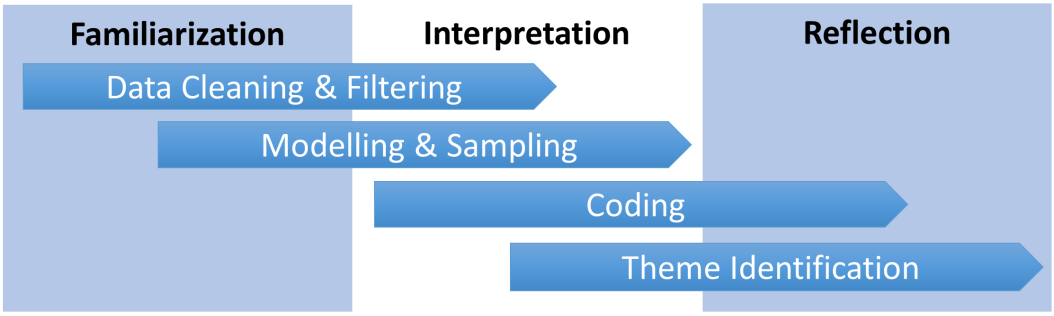


Fig. 1. Gauthier and Wallace [22]'s workflow that was used to focus the development of the toolkit. In moving from data collection to writing their final report, researchers progress through three conceptual stages of work: familiarization, interpretation, and reflection. To do so, they perform the practical tasks of data cleaning & filtering, modelling & sampling, coding, and theme identification. Like thematic analysis and computational methods, this workflow is highly iterative, and is not a linear process; researchers may shift between any conceptual stage or practical task.

- The **Modelling & Sampling module** enables researchers to identify and interpret latent patterns in the data using computational techniques, such as biterm topic modelling [50]. Once generated, researchers may label, merge, or remove topics, and visualize models as word lists and chord graphs. When a suitable model has been decided on, it may be used to purposively sample threads for further analysis.
- The **Coding module** enables researchers to manually code data, in an interface similar to qualitative data analysis software like NVIVO, MaxQDA, or Atlas TI. Researchers may choose from a list of sampled data, and develop a code tree by creating, modifying, and/or deleting codes.
- **Theme Identification** is supported through two modules. The **Reviewing module** enables researchers to create and review themes and codes through a network visualization. The **Reporting module** enables researchers to select and track the sources of quotations for the themes and codes developed in previous modules.

We deployed and iterated on version 0.8 of the Computational Thematic Analysis Toolkit [22]. The toolkit's full source code and installation files are available at https://osf.io/b72dm/

## 4 METHODS

To understand how the Computational Thematic Analysis Toolkit might influence research in practice, we performed a field comparison of two thematic analyses; one conducted through manual methods, and the other with computational methods (Figure 2). The two thematic analyses were performed independently, by expert teams, and with an inductive, realist perspective on the same set of real-world data. The manual team was asked to perform the analysis using their normal process. The computational team was asked to use the toolkit as they saw fit and told that it was not intended to replace their research, rather to provide a scaffold of tools to support their interaction with the data. Each team then reported on their process and findings, allowing us to compare and contrast each analysis and to develop an understanding of how the toolkit influenced the analysis.

The two analyses were conducted by teams of public health researchers, who were tasked with describing online discourses related to the generation and spread of rumours, misinformation and disinformation on COVID-19 in Canada. Both teams belonged to the same public health research group associated with the Canadian Immunization Research Network (CIRN), and were seeking the

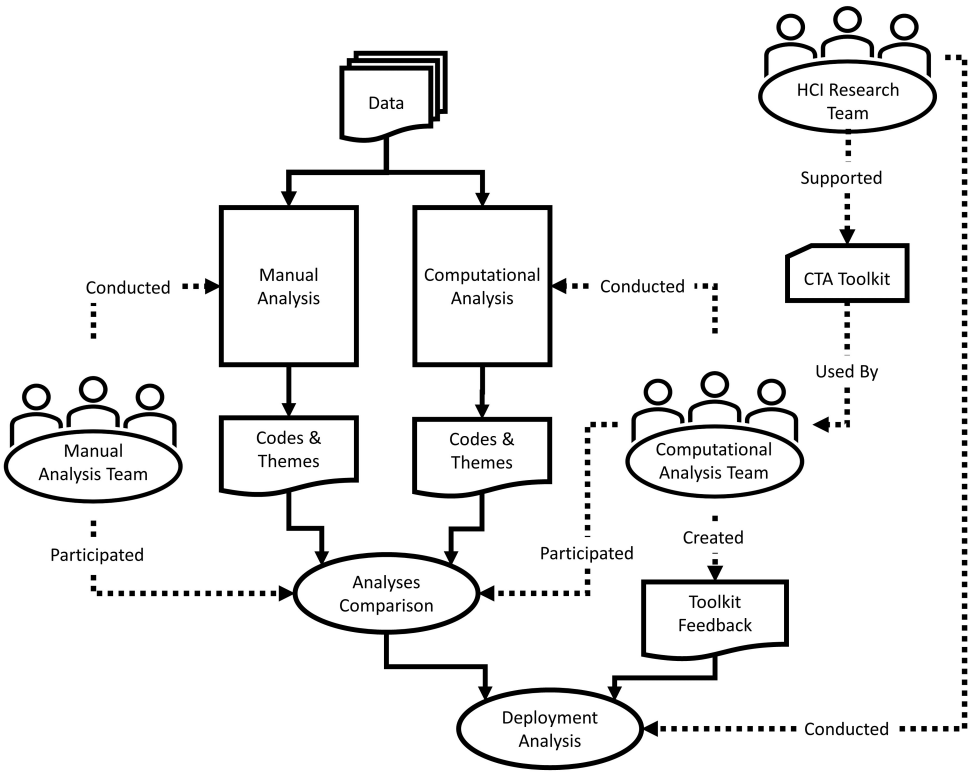


Fig. 2. Three research teams were involved in our case study. (1) The manual analysis team performed an inductive thematic analysis using their traditional methods. (2) The computational analysis team performed a thematic analysis using the Computational Thematic Analysis Toolkit [22]. (3) The HCI research team was responsible for technical support, and comparing the processes and outcomes of each thematic analysis.

ability to perform end-to-end analysis themselves after unsuccessfully working with an external team of AI consultants. Additionally, the public health research group: had not been involved in the toolkit's development; had research questions motivated by an emerging public health need; had expertise in conducting inductive thematic analyses; and had already collected data for their research questions.

Each team analyzed the same set of 613,666 English comments from Canadian news sites. This data set was collected by the CIRN team between January and June 2020 from a variety of Canadian news sites, including: CBC news, The Cape Breton Post, The Chronicle Herald, The Globe and Mail, The Halifax Examiner, The National Post, The Time Colonist, The Toronto Star, The Toronto Sun, The Tyee, and The Vancouver Sun. Thus, the members of both analysis teams were familiar with the data set.

### 4.1 Our Research Teams

Our research team comprised three sub-teams, each focused on a distinct task: toolkit development, computational analysis, and manual analysis. The members of both analysis teams came from a public health research group and had common experience conducting thematic analyses.

- The **manual analysis team** was responsible for using their research group's normal processes to conduct a thematic analysis of the shared data set and comparing the results of their manual analysis against the computational analysis. The team consisted of the third author, who is a trainee in anthropology, fourth author, who has a MSc in Public Health (Health Promotion and Program Evaluation) and investigates prevention, health promotion, communication, and misinformation, and sixth author, who has a PhD in public health and focuses on using qualitative methods to investigate the role of trust in Canadians' use of immunization and who was the principle investigator for funding acquisition. The third and fourth authors focused on developing and comparing results, while the sixth author assisted with preparing the data set, coordinating inter-team communications, and distributing results. Because of their inter-team communication role, the sixth author stayed at arms length during analysis tasks, such as coding, and comparison.
- The **computational analysis team** was responsible for conducting a thematic analysis using the Computational Thematic Analysis Toolkit, comparing the results of their computational analysis against the manual analysis, and providing feedback based on their experience. This team consisted of the second author, who has an MSc in Public Health and who researches prevention, vaccine hesitancy, social listening, and media coverage, and the fifth author, who has a PhD in medical anthropology and who specializes in using qualitative methods to investigate the socio-cultural field surrounding infectious diseases prevention. The second author acted as the primary analyst and the fifth author provided continual consultations throughout the analysis.
- The **HCI research team** provided technical support to the computational analysis team through training and bug fixes. They did not participate in the thematic analysis, but were responsible for analyzing the results of each and interpreting feedback from the other two teams. This team consisted of the first author, who is a PhD Candidate in public health and has software development, human computer interaction, and reflexive thematic analysis experience, and the seventh author, who is a professor and human-computer interaction researcher.

### 4.2 Data Collection & Analysis

To analyze the case study, we first gathered information about each analysis: (1) the processes, regarding what steps were performed and the activities that occurred during these steps; (2) the outputs, which consisted of themes and coding trees created during the analyses. We also collected (3) toolkit usage data from (a) the computational analysis team's journals, that described their thoughts as they used the toolkit; (b) a saved toolkit workspace, that captured the data used, the state of each module, and the actions that lead to these states; and (c) emails with the HCI research team, that describe activities in need of support during the analysis.

We then used the gathered data to describe both analyses' processes and outputs. To describe the analyses' processes, each analysis team summarized their notes taken while planning and conducting the analyses. These notes were then used by the HCI research team to create analysis process diagrams. In addition, for the computational analysis, the HCI research team added details of toolkit usage by triangulating the analysis' process with the toolkit usage. To describe the analyses' outputs, both analysis teams created coding trees, that capture the connections between themes and codes, as well as a description of each theme. These coding trees and the descriptions of themes were then integrated into tables by the HCI research team.

Finally, we compared the two analyses to identify similarities and differences. To compare the analysis, both analysis teams participated in a group discussion that went over both team's analyses' outputs and how these were created and used. For process, The HCI research team then triangulated

discussion notes with the descriptions of the two analyses to identify similarities and differences during both high level process steps and low level activities. For themes and codes created, the analysis teams summarized their discussions, which the HCI research team triangulated with the analyses' outputs to create a table of the overlap between topics.

## 5 RESULTS

In this section we present both the manual and the computational thematic analysis. First, we describe both analyses in terms of: (1) each team's analysis process and the activities performed; and (2) each team's results which are made up of themes and coding trees. We then identify similarities and differences by comparing the two processes and the team's results.

### 5.1 Manual Analysis

The manual analysis team followed a three step iterative process to analyze the 613,666 comments (Figure 3): (1) independent inductive coding, using 150 randomly selected comments; (2) group discussions, to establish and revise a coding framework; and (3) apply coding framework to a sample of 2,000 randomly selected comments. By the last iteration of this process the team created six main themes, one secondary theme, and a four level coding tree that links the codes to these themes (Table 1).

*5.1.1 Independent Inductive Coding.* The team first inductively coded a sample of 150 randomly selected comments. Two team members independently read the comments to develop context and then coded the comments to iteratively develop and identify initial codes, themes, and meanings that provided initial perspectives on the data set and how it could be analyzed. For instance, as a researcher read over the comments, they identified the general themes: criticism, frustration, and opposition from another user. As these general themes re-occurred, they created specified codes based on the subject matter of the comments, such as ageism, racism, stigma, discrimination, and xenophobia. The team completed an estimated 4 hours of coding the initial 150 comments on day 6. In the end the group of themes was combined into the theme Criticism with sub-themes corresponding to whom the comment was being critical towards, such as Towards User, Towards Experts, or Towards Governments.

*5.1.2 Group Discussions.* The team then initiated group discussions after (1) independent inductive coding and as an iterative activity when new codes were created during (3) applying coding framework. During these discussions, the team discussed how their codes apply to the data. The team held their initial discussion on day 6 for an estimated 4 hours and spent a further 2 hour of time meeting meeting to revise and finalize their coding framework over the course of step 3. For instance, they discussed examples from Criticism Towards User and whether they were identified as a Correction of facts to another user ensure that both coders were able to consistently distinguish between them. These discussions also helped the team to create and refine their common coding framework, which they used as a foundation for rest of the analysis. For instance, they discussed how Disinformation was an important theme during the pandemic on social media and how they could identify it, leading to the creation of codes for Minimization of the Virus, Conspiracy Theories, and Troll.

*5.1.3 Apply Coding Framework.* Finally, the team coded a sample of 2,000 randomly selected comments to assess code coverage, identify needed revisions, and themes. Two team members divided and deductively coded the comments using the common coding framework. Additionally, when the team identified the need for additional codes or revisions to existing codes they iterated back to (2) group discussions to revise their framework, and then resumed applying it to the sample.

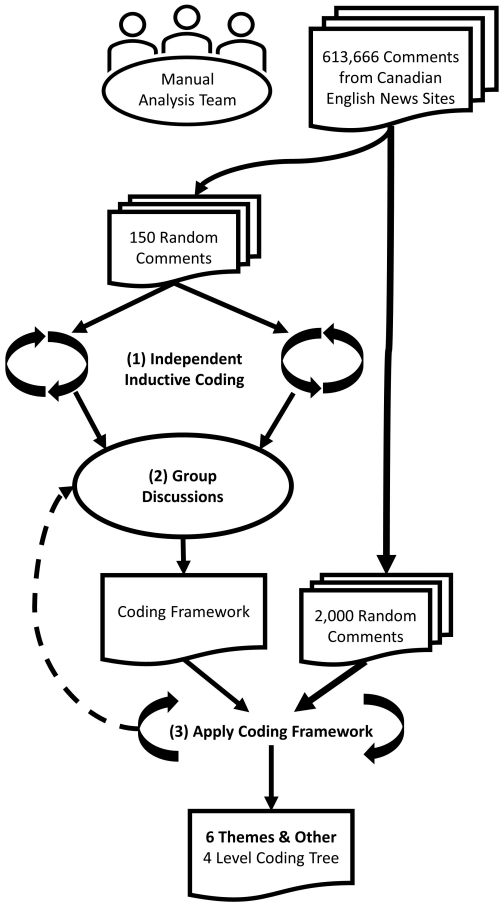


Fig. 3. The manual analysis team followed an iterative, three-step process : (1) Independent Inductive Coding, using a random sample of 150 comments; (2) Group Discussions, to establish and later revise a coding framework; and (3) Apply Coding Framework, using a random sample of 2,000 comments. The process created a four level coding tree that group codes using six main themes and one secondary theme.

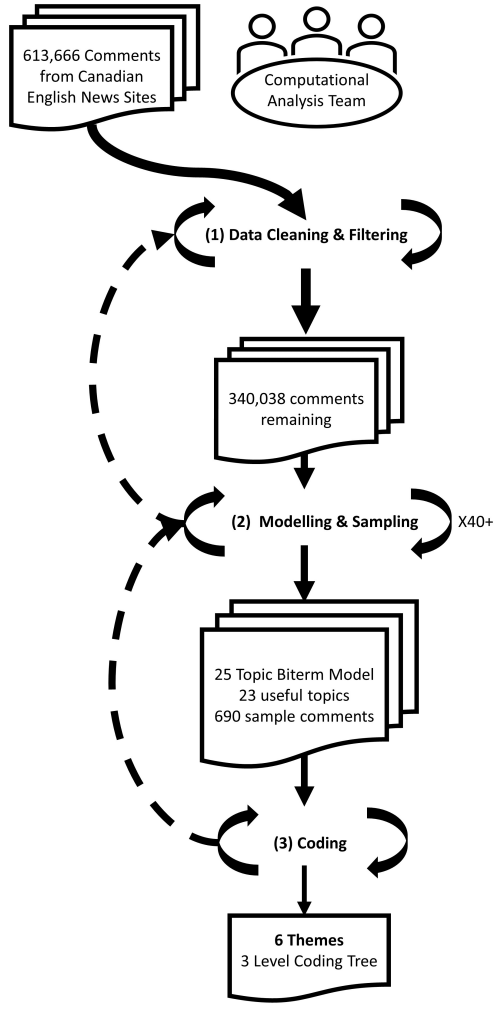


Fig. 4. The computational analysis team followed a separate, iterative, three-step process: (1) Data Filtering & Cleaning, which reduced the data from 613,666 to 340,038 comments; (2) Modelling & Sampling, which resulted a biterm topic model, and reduced the data down to a sample of 690 comments; and (3) Coding, during which they created a three level coding tree and developed six themes.

The team spent an estimated 14 hours on this step and finalized their application of codes and revising their coding tree on day 20. At the end of the process the they had developed a four level coding tree that grouped their codes under six main themes and one secondary theme (Table 1).

Table 1. Manual Analysis Coding Tree, made up of four levels. Six main themes and one Other theme make up the first level. Levels two, three, and four consist of codes that contributed to the themes.

1. **CRITICISM**
   Comments expressing an opinion on different aspects of the COVID-19 pandemic (sanitary measures, experts' opinions, authorities' decisions, media). This theme is divided into sub-themes, depending on who is targeted by the comment (another user, the experts, the population, the government or the media).
   - 1.1 **Towards User**
     - 1.1.1 Critical of those in favour of sanitary measures
     - 1.1.2 Critical of those in opposition to the sanitary measures/pandemic
     - 1.1.3 Correction of facts to another user
     - 1.1.4 Hateful comment
       - 1.1.4.1 Racism towards an individual
       - 1.1.4.2 Ageism
       - 1.1.4.3 Discrimination
       - 1.1.4.4 General insult
     - 1.1.5 Comment in agreement with another user
   - 1.2 **Towards Experts**
     - 1.2.1 Lack of trust in experts
     - 1.2.2 In agreement with the experts
   - 1.3 **Towards the Population**
     - 1.3.1 On non-compliance with sanitary measures
     - 1.3.2 Racism towards a group
     - 1.3.3 General criticism / discouragement
     - 1.3.4 encouragement
   - 1.4 **Towards Governments**
     - 1.4.1 USA / International
       - 1.4.1.1 In disagreement with the decisions
       - 1.4.1.2 Insult to Donald Trump
       - 1.4.1.3 In accordance with political decisions
     - 1.4.2 Canadian
       - 1.4.2.1 Poor budget management during the pandemic
       - 1.4.2.2 Poor management of the pandemic
       - 1.4.2.3 Good management of the pandemic
       - 1.4.2.4 Lack of confidence in the government
       - 1.4.2.5 In accordance with the government's decisions
       - 1.4.2.6 Insult to politicians
   - 1.5 **Towards Medias or the News**
     - 1.5.1 In agreement
     - 1.5.2 In disagreement
   - 1.6 **Towards Companies**

2. **DISEASE**
   Comments about the characteristics of COVID-19: transmission, origin of the virus, prevention and treatment, vaccines, statistics (cases/deaths), and screening.
   - 2.1 **Virus Transmission**
   - 2.2 **Screening - Testing**
   - 2.3 **Statistics/Deaths/Cases**
   - 2.4 **Origin of the Virus**
     - 2.4.1 Discrimination
     - 2.4.2 Racism
   - 2.5 **Prevention and Treatments**
   - 2.6 **Immunization/Vaccine**

3. **SANITARY MEASURES**
   Comments related to public health measures and recommendations. The theme is divided according to the view expressed by users (agree or disagree with the measure).
   - 3.1 **Confusion / Inconsistency of Measurements**
   - 3.2 **Skepticism About the Effectiveness of Measures**
   - 3.3 **In Accordance with the Measures**
     - 3.3.1 Wearing the mask
     - 3.3.2 Border closure
     - 3.3.3 Opening of the borders
     - 3.3.4 Physical distancing
     - 3.3.5 Lockdown and curfew
     - 3.3.6 Closing of non-essential businesses
     - 3.3.7 School and child care closures
     - 3.3.8 Opening of non-essential businesses
     - 3.3.9 Opening of schools and childcare services
   - 3.4 **Disagree with the Measures**
     - 3.4.1 Wearing the mask
     - 3.4.2 Border closure
     - 3.4.3 Opening of the borders
     - 3.4.4 Physical distancing
     - 3.4.5 Lockdown and curfew
     - 3.4.6 Closing of non-essential businesses
     - 3.4.7 School and child care closures
     - 3.4.8 Opening of non-essential businesses
     - 3.4.9 Opening of schools and childcare services

4. **IMPACTS OF THE PANDEMIC**
   Comments related to the impacts of the health measures and the management of the pandemic on different sectors of activity or social aspects (each represented by a sub-theme). The sub-themes are: work (health care workers, telecommuting, job loss, working conditions), societal impacts (school, long-term care, children, vulnerable populations), economic impacts and environmental impacts.
   - 4.1 **About the World of Work**
     - 4.1.1 Conditions of health care workers
     - 4.1.2 Remote work
     - 4.1.3 Loss of employment / return to the labour market
     - 4.1.4 Conditions of workers in general
   - 4.2 **Social Impact**
     - 4.2.1 Impact on the education system
     - 4.2.2 Propagation and death of the elderly / CHSLD
     - 4.2.3 Impacts on children
     - 4.2.4 Impacts on people with disabilities
     - 4.2.5 Impacts on Aboriginal communities
   - 4.3 **Economic Impacts**
   - 4.4 **Environmental Effects**

5. **DISINFORMATION**
   Comments related to conspiracy theories, downplaying the severity of the pandemic/virus and trolls.
   - 5.1 **Minimization of the Virus**
   - 5.2 **Conspiracy Theories**
   - 5.3 **Troll**

6. **INFORMATIVE COMMENT**
   Comments where informative content is shared and where users exchange information.
   - 6.1 **Argument/Information between Users**
   - 6.2 **Sharing References/Articles**

**OTHER**
Comments that are off-topic, not related to COVID-19 or not belonging to any of the previous themes.
- 1 **Out of Order**

### 5.2 Computational Analysis

The computational analysis team followed a three step process (Figure 4): (1) Data Filtering & Cleaning, to investigate words being used in comments and to focus what data computational techniques used; (2) Modelling & Sampling, to identify interesting patterns and generate useful samples of comments; and (3) Coding, using the samples to create their codes and themes. By the last iteration of this process they had created six themes and a three level coding tree that links the codes to these themes (Table 2). Throughout these three steps, the second and sixth authors spent an estimated five hours training and familiarizing themselves with the toolkit and four hours consulting with one another.

*5.2.1 Data Cleaning & Filtering.* The team first iteratively inspected, cleaned, and filtered the words being included and removed. To make these choices, they interpreted the lists of words through the lens of their knowledge of the vaccine discourse topic and data set to determine relevance to COVID-19 commentary. When they encountered unfamiliar words, they looked at comments that used the word to expand their understanding. Words were excluded that either occurred fewer than 20 times or were considered to be 'noise', such as conjunctions, interjections (e.g., 'oh', 'ah'), insults, and misspelled words. They also removed the names that were not connected to public figures of interest.

Over the course of their analysis, the team created a contextually useful set of 578 filtering and cleaning rules that reduced the data down to 340,058 comments. They used these rules to create a total of 46 models, exploring different rules' impact on the words in the data. The team spent an estimated 8 hours iteratively adding rules and inspecting included and removed words over the first 15 days of the analysis.

*5.2.2 Modelling & Sampling.* The team then iteratively constructed models to interpret and sample data for thematic analysis. They configured model-specific parameters, such as number of topics and passes for biterm, and then inspected the generated models, and removed, merged, and labelled topics, selected comments for later coding, and considered potential codes. After learning from each model, they returned to either the first step, to perform additional word filtering and cleaning, or to build new models with adjusted parameters (such as number of topics). The team completed generating models after 15 days and an estimated 15 hours.

After more than 46 iterations, the team selected model 41 as a foundation for their coding activities (Figure 5). Although they generated model 41 on day 14, the team considered it more contextually valuable than models from further iterations on days 14 and 15. Model 41 was a biterm model generated with 25 topics and 500 passes and the team felt that its topics were cohesive and useful for the context of their vaccine hesitancy research area. After they had inspected and interpreted the model, 23 topics remained that were useful for identifying comments for the third step.

*5.2.3 Coding.* Finally, the team created and applied codes to develop themes and a coding tree. First, they used model 41 to sample 30 comments from each of its 23 topics, for a total of 690 comments. They then inductively coded the comments by iteratively selecting and interpreting each comment. After completing all coding iterations, they had created and applied 38 codes. Using these codes, they created a document that describes a three level coding tree that grouped the codes under six themes (Table 2). The team spent an estimated 17 hours coding and finished their analysis on day 25 of the analysis.

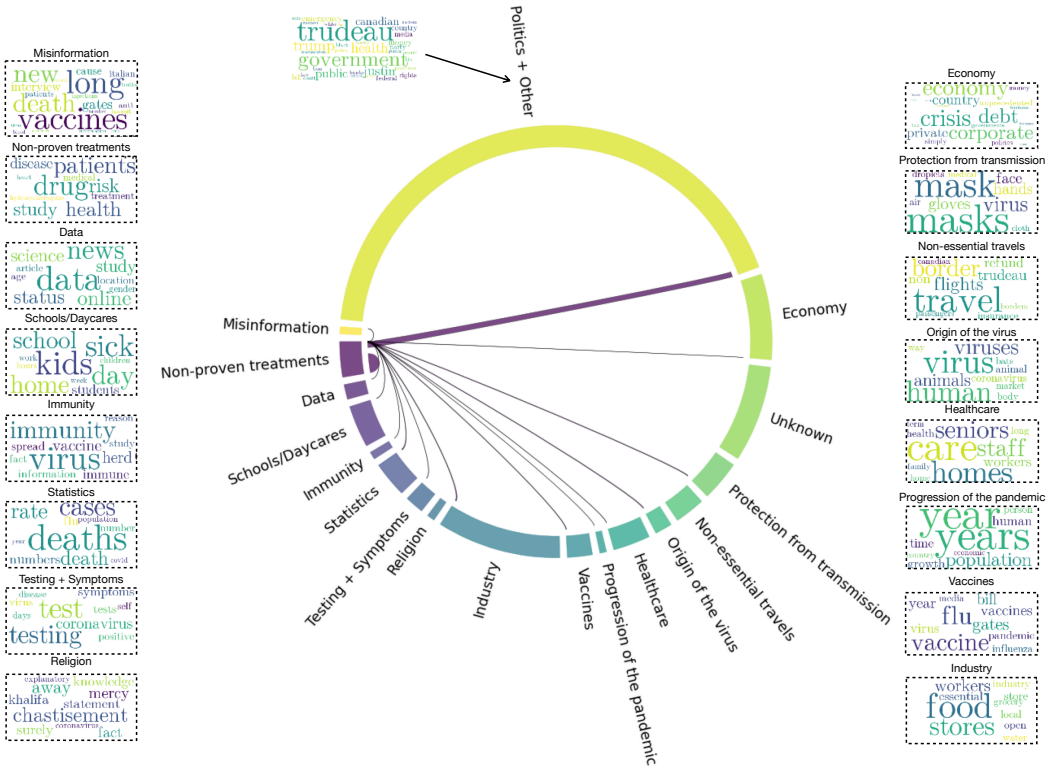


Fig. 5. A chord diagram of Model 41, which the computational analysis team used to derive their code tree. Each topic show the human created labels. Within the Computational Thematic Analysis toolkit, these chord diagrams are interactive: topic labels can be clicked on to show/hide word clouds of topic keywords, and mousing over arcs for each topic shows chords that represent document overlap.

### 5.3 Comparison of Processes and Outcomes

Both analysis teams followed an iterative process and performed inductive coding throughout their analyses. However, when comparing their workflows and outcomes, we identified some (sometimes subtle) differences between the two groups. We now discuss these differences in terms of toolkit usage, differences in time, differences in process, and similarities and differences in outcomes.

*5.3.1 Toolkit Usage.* The CTA toolkit enabled qualitative researchers to learn applied skills, such as data cleaning and adjusting model parameters, and to tune and generate useful models. Such applied skills are needed when integrating existing computational techniques in a qualitative research process, as opposed to programming skills that are required to implement and refine computational techniques into usable tools. Additionally, this learning helped the researchers develop their contextual knowledge of how and when different techniques can be integrated with specific thematic analysis tasks. For instance, the computational analysis team chose to integrate the biterm topic model as a primer for coding, based on it grouping common word patterns, rather than as a replacement for the entire theme and coding process.

Alongside the applied skills, the computational analysis team developed contextual understandings of limitations of computational techniques as they used the CTA Toolkit. For instance, the

Table 2. The coding tree generated by the computational analysis team. It was a three-level tree, with six themes present in the first level. Levels two and three consist of codes that contributed to each theme. Codes that originated from interpreting topics from a model during step 2 are denoted by ★. Codes that originated from coding comments during step 3 are denoted by †. Other codes and themes were developed when organizing the coding tree.

1. **USER SHARING OPINIONS**
   Comments expressing an opinion. It is divided into 3 subtopics: opinion on the management of the pandemic by the authorities, criticism of the media and criticism of people not complying with preventive measures.
   - 1.1 Opinion on the Management of the Pandemic†
   - 1.2 Criticism of Media†
   - 1.3 Criticism of People Not Complying with Preventive Measures†

2. **DISEASE**
   Comments related to the characteristics of COVID-19: origin of the virus, transmission, symptoms, perceived risk, methods of protection against transmission, testing, immunity, and progression of the virus.
   - 2.1 **Origin of the Virus★**
     - 2.1.1 China
     - 2.1.2 Animals
     - 2.1.3 Made by human
     - 2.1.4 Act of God (Religion)★
   - 2.2 Protection from Transmission★
   - 2.3 Testing★
   - 2.4 Immunity★
   - 2.5 Progression of the pandemic★
   - 2.6 Transmission†
   - 2.7 Symptoms†
   - 2.8 Risk†

3. **PREVENTION AND TREATMENT**
   Comments on ways to prevent or treat COVID-19, including unproven or alternative treatments and vaccines.
   - 3.1 Vaccines★
   - 3.2 Non-Proven Treatments★

4. **IMPACTS OF THE PANDEMIC**
   Comments about the impacts of health measures and pandemic management on different sectors of activity or social aspects (each represented by a sub-theme). The sub-themes are: school/daycare, industry and business, interpersonal relationships, travel industry, health system, economy and stigma/racism.
   - 4.1 Impacts of COVID-19 on Schools/Daycares★
   - 4.2 Impacts of COVID-19 on Industries and Businesses★
   - 4.3 Impacts on Interpersonal Relationships†
   - 4.4 **Non-Essential Travel★**
     - 4.4.1 Impacts on travellers and airlines
     - 4.4.2 Preventive measures for travellers
     - 4.4.3 Border closure
   - 4.5 **Healthcare★**
     - 4.5.1 Long term care
     - 4.5.2 Healthcare workers
     - 4.5.3 Equipment
   - 4.6 **Economy★**
     - 4.6.1 Financial aid
     - 4.6.2 Economic crisis/debt
   - 4.7 Stigma and Racism†

5. **TYPES OF INFORMATION**
   Comments where informational content is shared. It is divided into 2 subtopics: statistics/data on COVID-19 (cases, deaths) and misinformation.
   - 5.1 Statistics & Data★
   - 5.2 Misinformation★

6. **NOT NECESSARILY RELATED TO COVID-19**
   Comments related to politics (provincial, national and international). At the time of data collection, the political context was particular, both in Canada and in the United States (tensions between the Conservatives and the Liberals, Donald Trump). Although efforts were made to eliminate purely partisan comments during the data cleaning and filtering process, many remain and it is difficult to determine whether or not they are related to the pandemic. Finally, comments posted by trolls and off-topic comments are also classified under this theme.
   - 6.1 Politics†
   - 6.2 Troll/People Insulting Each Other†
   - 6.3 Other†

team became aware that frequently repeated data, such as copy and pasted comments, can limit modelling techniques and distort model results, making it important to manage through both filtering and interpretation. Understanding of such limitations helps qualitative researchers avoid over-reliance on computational techniques which, while useful tools to interact with data, are not replacements for domain expertise or a human researcher's interpretive abilities.

*5.3.2 Differences in Time.* The manual analysis team completed approximately 24 hours of work over 25 days. They estimated that work was distributed between several analytical sub-tasks:

- ~4 hours Coding their initial data set of 150 comments
- ~6 hours Group Discussions
- ~14 hours Applying their coding framework to 2000 random comments

On the other hand, the computational analysis team reported working for approximately 49 hours, spread over a 30-day period in which they were also working on several ongoing, unrelated projects. The team provided further estimates for analytical sub-tasks:

- ~5 hours Initial orientation with the toolkit
- ~8 hours Data Cleaning & Filtering
- ~15 hours Modelling & Sampling
- ~17 hours Coding
- ~4 hours Participating in team consultations

The team also reported that some of this time was spent exploring different software features (e.g., token filtering and topic modelling) and that in some cases the toolkit was less mature than the commercial software they typically use (i.e., NVIVO [30]). In particular, the CTA toolkit's coding interface lacked features such as multi-comment coding and visualizations for code hierarchy, was less polished than commercial software, and was delivered to the team with a few performance-impacting bugs that were quickly addressed by the development team, but which slowed the analysis team down.

*5.3.3 Differences in Process: Top-Down vs Bottom-Up.* The computational analysis team used a top-down process where they: (1) followed the toolkit's large-to-small data handling approach that reduced the amount of data from 613,666 comments down to 690 comments; (2) aligned their steps to match the toolkit's modules and integrated human activities into each step (Figure 4); and (3) started coding by covering multiple comments and then revising their coding to capture specific comments' context. This top-down process jump-started the team's interpretation of recurring patterns in the data *before* creating the codes. The toolkit's word list visualizations (Data Filtering & Cleaning) helped the team identify recurring topic keywords from across the data like 'Trudeau', 'masks', 'money', and 'businesses'. They then looked for reoccurring keywords across different models (Modelling & Sampling) to identify topics like Immunity, Vaccines, Statistics, and Protection. The team's interpretation of these recurring topics and keywords were the origin points of 15 codes in their final tree, such as 2.2 Protection from Transmission (denoted by ⋆ in Table 2).

On the other hand, the manual analysis team used a bottom-up process where they: (1) followed a small to large data handling approach, that included 150 comments at first and added 2000 comments; (2) defined their steps by the human analysis activities (Figure 3); and (3) started their coding by capturing specific comments' contexts and then refining the coding to cover multiple comments. They relied on developing specific codes for individual comments and multiple encounters before grouping recurrences. For example, they created nine subcodes to capture comment-specific sanitary measures before grouping them into two mirrored parent codes, 3.3 In Accordance with the Measures and 3.4 Disagree with the Measures, that specify two different recurring contexts).

These two approaches lead to subtle differences in the teams' coding processes. First, each team interpreted codes in different ways. For instance, in the context of mask use, the computational analysis team considered the discussion of masks sufficient to assess whether the code 2.2 Protection from Transmission occurred. On the other hand, the manual analysis team needed to interpret whether each comment's context aligned with agreeing or disagreeing with masks to fit one of the sub-codes 3.3 In Accordance with the Measures or 3.4 Disagree with the Measures. Second, the computational analysis team maintained their coding focus by grouping potentially similar comments based on model 41's interpreted topics group to provide structure. In comparison, the manual analysis team had to reset focus between different comment types encountered during coding until they developed an internal interpretation of similar comment groupings. During the comparison, the teams decided that the computation analysis coding process was easier to perform while the manual analysis coding supplied more specific details about the data.

Table 3. Overlap between manual and computational analysis themes.

| | | Manual Analysis Themes | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Criticism | Disease | Sanitary Measures | Impacts of the Pandemic | Disinformation | Informative Comment | Other |
| Computational Analysis Themes | Users Sharing Opinions | ● | | | | | | |
| | Disease | | ● | ● | | | | |
| | Prevention and Treatment | | | ● | | | | |
| | Impacts of the Pandemic | | | | ● | | | |
| | Types of Information | | | | | ● | ● | |
| | Not Necessarily Related to COVID-19 | ● | | | | ● | | ● |

*5.3.4 Similarities in Code Trees.* Both analysis teams created coding trees that organized the inductive codes, developed during their interactions with the data. In these structures the themes were the first level and any additional level consisted of codes connected to the theme. Both teams used their coding trees for two purposes: to consistently apply their codes across the data; and to ground communication when discussing themes and codes, both within the teams during their analyses and across teams when comparing the analyses.

In both coding trees the themes were positioned as parent nodes and overlapped by covering the same ideas despite having slightly different names and descriptions (Table 3). The teams stated in their comparison summary that these differences were expected and perfectly normal for two independent inductive thematic analyses, as developing the themes from data involves subjective description of ideas. Based on the similarities between the two code trees, the teams felt that the toolkit had helped the computational analysis team develop real themes.

*5.3.5 Differences in Code Trees.* However, when debriefing, the analysis teams also acknowledged that different coding tree structures had been created and used. The computational analysis' coding tree had three levels, was focused on general descriptions, and was useful for capturing common ideas from multiple comments under each code which made coding large numbers of comments easier. The manual analysis' coding tree had four levels, was focused on specialized descriptions, and was useful for describing the specific comments coded which made it time consuming to apply to large numbers of comments. Despite having different structures, the teams agreed that neither coding tree was invalid, rather the two trees provided distinct forms of utility and both could contribute to a successful thematic analysis.

We also identified a difference in how the coding trees were reported. The computational analysis team included indicators of where codes originated from, either topic model interpretation, coding, or code organization. This additional information provided transparency about how codes may connect to the interpretation of computational techniques. In comparison, the manual analysis team did not indicate code origin in their coding tree as all codes originated from human inductive coding.

## 6 DISCUSSION: AGENCY AND AMPLIFICATION

We found that the CTA Toolkit provided access to advanced computational techniques for researchers without programming expertise, facilitated their interaction and interpretation of the data, but also that it influenced how they approached their thematic analysis. We now discuss how each of these findings can inform ongoing efforts by the human-computer interaction and machine learning research communities to integrate computational techniques into qualitative methods, like thematic analysis.

### 6.1 Opening Computational Techniques to Non-Programmers

The CTA Toolkit served as a scaffold upon which researchers could use computational techniques to develop interpretations of data [6]. By the end of their analysis, the team had tried more than 40+ combinations of model parameters and types. Similarly, the team explored NLP summary data, iteratively filtered words based on their context and usage, and observed the impact of that filtering on the generated models. The non-programmer, qualitative researchers used these computational techniques on their own, rather than through a third party such as an AI consultant who may not have the same understanding of the research area or objectives.

Several members of the human-computer interaction research community have raised the question of 'perfect' being the enemy of 'good' [6, 21]. For instance, one might question whether these analyses were performed 'optimally', and whether the same tools may have yielded more accurate models in the hands of experts. However, our post-analysis comparison of themes and coding trees shows that both the computational and manual teams created similar results. Qualitative researchers could also use what they learned from the models to describe their choices, which contributes to establishing process transparency and the trustworthiness of analysis results [41, 46, 47].

Taken together, these activities demonstrate how tools can lower the threshold to use computational techniques during qualitative research [5, 26], establish common ground between computation and qualitative communities [17], and create space where further collaborations can benefit both fields [5, 17]. And so we suggest that with tools like the CTA toolkit there is a real opportunity to start applying computational techniques in practice, to better understand the actual challenges qualitative researchers face when performing computationally-supported research, and to refocus the human-computer interaction research community on solving them. That is, there is an opportunity to engage in truly human-centred data science.

### 6.2 Facilitating Interaction and Aiding Interpretation

The CTA toolkit also 'raised the ceiling' for thematic analysis of large, online data sets [29]. In follow-up discussions, the computational analysis team reported that the CTA Toolkit enabled them to focus their efforts on data interpretation. Iterations of filtering and modelling tasks enabled them to shift from a sample of 613,666 comments, to 340,038 comments, to an informative sample of 690 comments (i.e., purposive and judgmental sampling [25, 33]). The same models made it easier to locate topics of interest and groups of comments to code. The team also described how these computational tasks helped them to create meaning [17]. For instance, the Healthcare code (Table 2, 4.5) and its subcodes for Longterm care, Healthcare workers, and Equipment were interpreted from interacting with a topic model and its keywords like care, homes, seniors, staff, workers, health, long, term, home, family, and masks.

However, the CTA toolkit did not facilitate all of the analysis team's tasks. For instance, although it helped them clean and filter the data, removal or inclusion of words was too limited to manage some contextual signals and/or noise. As one example, synonyms for Canadian Prime Minister

'Justin Trudeau' like 'trudeau', 'trudeaus', 'trudo', and 'justintrudeau' all contributed to codes (Table 2, Opinion on the Management of the Pandemic and Politics). But this work had to be performed manually because the toolkit did not support finding and replacing synonyms. Similarly, comments were sometimes difficult to interpret when they referenced ideas that were specific to a subset of the data, such as a specific news organization, which topic-based sampling does not account for.

These limitations help to highlight that the ceiling can be raised, and to reveal how computational techniques can further amplify qualitative researchers' ability to interpret data. Data cleaning and filtering can be expanded to align with researchers' interpretative activities, such as enabling researchers to manipulate the data to represent their interpretation of the synonyms. Similarly, model generation and visualizations could occur more directly, and *collaboratively*, to facilitate researchers' interpretation (e.g., [19]). Coding and theme identification activities might further be supported by tools that help researchers reflect and reconsider relationships between data and their interpretations of it.

### 6.3 Influence of Computational Techniques on Research Process

Importantly, the computational analysis team felt that they maintained autonomy, intimacy, and ownership of their analysis [26]. Indeed, they reported that the CTA toolkit assisted with specific analysis tasks, particularly pattern identification and sampling [20]. We attribute this sense of control to the toolkit's design, which was intended to be flexible and support researchers' own styles and preferences [22]. Researchers have substantial agency in terms of choosing how and when to iterate, determining which models are useful, and quickly experimenting with different filters and models in a safe environment. The CTA toolkit made common data science tasks rapid, incremental, and reversible — key principles for intuitive and predictive interfaces [28] — and, when tool limitations were encountered, researchers were able to fall back on their expertise and practical knowledge.

These findings demonstrate the potential of visual interfaces to enable non-programmers to use computational methods in their analysis. But our comparison of the teams' processes and outcomes also points to some previously unidentified side-effects. In particular, we observed differences in how each team sampled comments for inspection and coding, and how those codes ultimately translated into code tree artifacts.

First, while the two code trees produced by each team were similar, they arrived at them using quite different processes. This divergence was not a surprise — the CTA toolkit was designed to enable model-based sampling and bag-of-words pattern identification that is simply not feasible without the aid of computational techniques. However, it's currently unclear under which conditions one might prefer the top-down vs. bottom-up approaches taken by our two teams; are there circumstances under which a team would favour one over the other? Can computational techniques support a bottom-up analysis? Or perhaps they should be considered in a 'hybrid' fashion, similar to what was proposed by Muller et al. [36]. We leave these questions to future work.

Second, the trees themselves differed in the amount of detail present, and their use in supporting the thematic analysis. The code tree developed by the manual analysis team was more detailed, and was used throughout the process to foster common ground — at least in some sense, as a means to coordinate team members' coding activity (i.e., inter-coder reliability [34]). On the other hand, the computational analysis team used the CTA Toolkit together and thus did not need to use the coding tree to develop internal common ground. The long-term implications of these differences are unclear: might they impact a team's ability to share, expand on, or otherwise further develop their results? Might they impact transferability or transparency?

These differences may be more insidious than the adoption concerns raised by Jiang et al. [26], since they were not consciously raised or considered by researchers during the analysis, but were apparent in post-hoc comparisons of process and outputs. They also complicate calls to action, particularly those from Braun et al. [15] and Braun and Clarke [14], in supporting transparent reporting and rigour in qualitative research. That is, if researchers are unaware of how computational methods might influence their research, how can they adequately reflect and report on those influences?

### 6.4 Next Steps in Critical Computing Research

Finally, our case study provides an opportunity for the human-computer interaction and critical computing research communities to consider next steps in advancing computationally-supported qualitative research. Our research provides an opportunity to triangulate with contemporary findings like Jiang et al. [26] and Feuston and Brubaker [20] that interviewed qualitative researchers about their experiences, aspirations, and concerns. It also provides an opportunity for calls-to-action for cross-pollination between human-computer interaction and qualitative research communities: human-computer interaction researchers need to better understand qualitative methods [5, 17, 20, 26], and qualitative researchers need to better understand computational methods [2, 16, 39, 44, 48].

Much of the contemporary discourse (e.g., [16, 44, 48]) still considers data science in a segregated context, where technical scientists with programming experience independently develop models before deploying them to a team of domain experts, stakeholders, or 'users' [12]. This segregation places an emphasis on 'eager AI' [20] that then needs to be *interpretable* or *explainable*. It also emphasizes the mathematical optimality and pseudo-objectivity of models over their usefulness to domain experts, which the human-computer interaction research community has sought to address through concepts like computational reflexivity and model positionality [16]. Given this segregation it is hardly surprising that, when asked, qualitative researchers raise concerns about about loss of agency in their analysis process [20, 26].

Our work contributes to a growing body of research that seeks to more fully engage qualitative researchers in the design and use of ML and AI. To date, that research has primarily relied upon the development of prototype tools (e.g., [7]) and interviews with domain experts (e.g., [20, 26]). Our case study further builds on existing calls 'put tools in their place' [20], and to consider how ML and AI can empower domain experts without requiring them to become experts in computer programming. We stress the need for additional research methods — particularly those that emphasize realism [35] like case studies, autobiographical design, and research through design [38, 51] — to further our understanding of ML workflows.

## 7 LIMITATIONS

In this work, we conducted a study of two thematic analyses performed 'in the wild'. That is, two research teams with expert knowledge in public health engaged with a large data set obtained from Canadian media to answer pressing questions about vaccine hesitancy during the SARS-CoV-2 pandemic. Conducting such a case study provides a rich environment in which to explore a highly subjective and interpretive research process with an emphasis on realism at the expense of experimental precision and control [35]. And so, there are some inherent limitations and benefits to this approach:

First, a limitation of exploratory human-computer interaction research is that identifying appropriate participants is often a challenge, or even impossible, since such a target audience may not yet exist [29]. While qualitative research, and reflexive thematic analysis [13, 15] in particular, are extremely popular, valuable research strategies one might not consider our research team to be expert 'end users' for the Computational Thematic Analysis Toolkit [22]. That is, this was the

research team's first analysis using the toolkit, and one would expect practices to evolve as they became more adept with it and gained experience with different data sets, analyses, etc.

Second, thematic analysis is an interpretive process [13, 15], and interpreting and comparing research outputs between two teams is extremely complex. It is difficult to argue that any one code tree is 'better' than another. One also needs to consider what was learned during the analysis process itself, and there are many potential trade-offs between time spent on analysis and the researcher's depth of understanding, whether a researcher choose to explore themes in depth or breadth, and which perspectives they chose to explore and emphasize. In our case study, many internal and external factors also influenced the amount of time it took each team to perform their analysis, including: the time required by the computational team to familiarize themselves with the toolkit, differences in their research approach (e.g., single vs multiple coders), and real-world stressors like family commitments and other concurrent projects. All of these factors were beyond our control. Instead, we focused our analysis on the *research process*, and understanding how the manual and computational methods influenced it.

## 8 CONCLUSION

In this work we explored how the integration of computational techniques into thematic analysis plays out in real world research. Our teams conducted two independent thematic analyses of 613,666 online communities, one manual and one computational using the CTA Toolkit. We presented results that describe both analyses' process and outputs, and then compared them to identify their similarities and differences.

Grounded in this case study, we identified the benefits and opportunities of using computational techniques to augment qualitative analysis of large data sets. We showed how the Computational Thematic Analysis toolkit made computational techniques accessible to non-programmer researchers, and enhanced their ability to interpret large data sets. We also found that researchers maintained a sense of agency during the analysis, contrary to concerns raised in previous research, but showed how the tool subtly influenced how researchers approached their analysis. In discussing these findings, we shared provocations that future research should explore how to avoid these (potentially) insidious influences of computational methods, and incorporate real-world deployments of technology to understand how they play out in practice.

## 9 ACKNOWLEDGEMENTS

We thank the University of Waterloo's School of Public Health Sciences, Centre de recherche du CHU de Québec-Laval University, and HCI + Health Lab for providing constructive and supportive communities. This work was made possible by NSERC Discovery Grant 2015-06585 and Canadian Immunization Research Network Grant FRN#151944.

## REFERENCES

[1] Salim Ahmed, Rebecca J. Haines-Saah, Arfan R. Afzal, Helen Tam-Tham, Mohammad Al Mamun, Brenda R. Hemmelgarn, and Tanvir C. Turin. 2017. User-driven conversations about dialysis through Facebook: A qualitative thematic analysis. *Nephrology* 22, 4 (2017), 301–307. https://doi.org/10.1111/nep.12780

[2] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human-Centered Data Science: An Introduction.* MIT Press, Cambridge, Massachusetts.

[3] Jodi Aronson. 1995. A Pragmatic View of Thematic Analysis. *The Qualitative Report* 2, 1 (Jan 1995), 1–3.

[4] Angelica Attard and Neil S. Coulson. 2012. A thematic analysis of patient communication in Parkinson's disease online support group discussion forums. *Computers in Human Behavior* 28, 2 (Mar 2012), 500–506. https://doi.org/10.1016/j.chb.2011.10.022

[5] Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2021. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures* 0, 0 (Dec 2021), 1–18. https://doi.org/10.1080/19312458.2021.2015574

[6] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410. https://doi.org/10.1002/asi.23786 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23786

[7] Eric P. S. Baumer, Drew Siedel, Lena McDonnell, Jiayun Zhong, Patricia Sittikul, and Micki McGee. 2020. Topicalizer: reframing core concepts in machine learning visualization by co-designing for interpretivist scholarship. *Human–Computer Interaction* 35, 5 (2020), 452–480. https://doi.org/10.1080/07370024.2020.1734460 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2020.1734460.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022. http://www.jmlr.org/papers/v3/blei03a.html

[9] Marcela C. C. Bomfim, Sharon I. Kirkpatrick, Lennart E. Nacke, and James R. Wallace. 2020. *Food Literacy While Shopping: Motivating Informed Food Purchasing Behaviour with a Situated Gameful App.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376801

[10] Michael Bostock and Jeffrey Heer. 2009. Protovis: A Graphical Toolkit for Visualization. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov 2009), 1121–1128. https://doi.org/10.1109/TVCG.2009.174

[11] Richard E. Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development.* Sage Publications, Inc, Thousand Oaks, CA, USA.

[12] Adam Bradley, Cayley MacArthur, Mark Hancock, and Sheelagh Carpendale. 2015. Gendered or neutral? Considering the language of HCI. In *Proceedings of the 41st graphics interface conference.* ACM, New York, New York, USA, 163–170.

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[14] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (Aug 2021), 1–25. https://doi.org/10.1080/14780887.2020.1769238

[15] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. *Thematic Analysis.* Springer, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[16] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3501998

[17] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.* 8, 2 (Jun 2018), 9:1–9:20. https://doi.org/10.1145/3185515

[18] Alexandra R. D'Agostino, Allison R. Optican, Shaina J. Sowles, Melissa J. Krauss, Kiriam Escobar Lee, and Patricia A. Cavazos-Rehg. 2017. Social networking online to recover from opioid use disorder: A study of community interactions. *Drug and Alcohol Dependence* 181 (Dec. 2017), 5–10. https://doi.org/10.1016/j.drugalcdep.2017.09.010

[19] Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. 2019. Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 374–384. https://doi.org/10.1109/TVCG.2018.2864769

[20] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct 2021), 469:1–469:25. https://doi.org/10.1145/3479856

[21] Robert P Gauthier, Mary Jean Costello, and James R Wallace. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 20, 17 pages. https://doi.org/10.1145/3491102.3502076

[22] Robert P. Gauthier and James R. Wallace. 2022. The Computational Thematic Analysis Toolkit. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 25 (jan 2022), 15 pages. https://doi.org/10.1145/3492844

[23] Rebecca J. Gooden and Helen R. Winefield. 2007. Breast and Prostate Cancer Online Discussion Boards: A Thematic Analysis of Gender Differences and Similarities. *Journal of Health Psychology* 12, 1 (Jan. 2007), 103–114. https://doi.org/10.1177/1359105307071744 Publisher: SAGE Publications Ltd.

[24] Jason Hill and Carl Gutwin. 2004. The MAUI Toolkit: Groupware Widgets for Group Awareness. *Computer Supported Cooperative Work (CSCW)* 13, 5 (Dec 2004), 539–571. https://doi.org/10.1007/s10606-004-5063-7

[25] Orland Hoeber, Larena Hoeber, Ryan Snelgrove, and Laura Wood. 2017. Interactively Producing Purposive Samples for Qualitative Research using Exploratory Search.. In *SCST@ CHIIR.* CEUR-WS, Oslo, Norway, 18–20.

[26] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (Apr 2021), 1–23. https://doi.org/10.1145/3449168 arXiv: 2102.03702.

[27] Tejinder K. Judge, Carman Neustaedter, and Andrew F. Kurtz. 2010. The Family Window: The Design and Evaluation of a Domestic Media Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2361–2370. https://doi.org/10.1145/1753326.1753682

[28] Bum chul Kwon, Waqas Javed, Niklas Elmqvist, and Ji Soo Yi. 2011. Direct Manipulation Through Surrogate Objects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 627–636. https://doi.org/10.1145/1978942.1979033

[29] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation Strategies for HCI Toolkit Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, Montreal QC, Canada, 1–17. https://doi.org/10.1145/3173574.3173610

[30] QSR International Pty Ltd. 2021. NVivo qualitative data analysis software. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

[31] Blair MacIntyre, Maribeth Gandy, Steven Dow, and Jay David Bolter. 2004. DART: a toolkit for rapid design exploration of augmented reality experiences. In *Proceedings of the 17th annual ACM symposium on User interface software and technology (UIST '04)*. Association for Computing Machinery, New York, NY, USA, 197–206. https://doi.org/10.1145/1029632.1029669

[32] Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. 2018. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures* 12, 2-3 (2018), 93–118. https://doi.org/10.1080/19312458.2018.1430754 arXiv:https://doi.org/10.1080/19312458.2018.1430754

[33] Martin N. Marshall. 1996. Sampling for qualitative research. *Family Practice* 13, 6 (Jan 1996), 522–526. https://doi.org/10.1093/fampra/13.6.522

[34] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (nov 2019), 23 pages. https://doi.org/10.1145/3359174

[35] Joseph E. Mcgrath. 1995. Methodology Matter: Doing Research in the Behavioral and Social Sciences. In *Readings in Human–Computer Interaction*, Ronald M. Baecker, Jonathan Grudin, William A. S. Buxton, and Saul Greenberg (Eds.). Morgan Kaufmann, USA, 152–169. https://doi.org/10.1016/B978-0-08-051574-8.50019-4

[36] Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. ACM, New York, NY, USA, 3–8. https://doi.org/10.1145/2957276.2957280

[37] Brad Myers, Scott E. Hudson, and Randy Pausch. 2000. Past, present, and future of user interface software tools. *ACM Transactions on Computer-Human Interaction* 7, 1 (Mar 2000), 3–28. https://doi.org/10.1145/344949.344959

[38] Carman Neustaedter and Phoebe Sengers. 2012. Autobiographical design in HCI research: designing and learning through use-it-yourself. In *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. Association for Computing Machinery, Newcastle Upon Tyne, United Kingdom, 514–523. https://doi.org/10.1145/2317956.2318034

[39] Orestis Papakyriakopoulos, Elizabeth Anne Watkins, Amy Winecoff, Klaudia Jaźwińska, and Tithi Chattopadhyay. 2021. Qualitative Analysis for Human Centered AI. Method Poster from Human Centered AI workshop at Neural Information Processing Systems 2021. , 8 pages. https://doi.org/10.48550/arXiv.2112.03784

[40] Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. Alto: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the (Volume 1: Long Papers)*, Vol. 1. Association for Computational Linguistics, Baltimore, MA, USA, 1158–1169. Issue 1.

[41] Michael G. Pratt, Sarah Kaplan, and Richard Whittington. 2020. Editorial Essay: The Tumult over Transparency: Decoupling Transparency from Replication in Establishing Trustworthy Qualitative Research. *Administrative Science Quarterly* 65, 1 (Mar 2020), 1–19. https://doi.org/10.1177/0001839219887663

[42] Shelly Rodgers and Qimei Chen. 2005. Internet Community Group Participation: Psychosocial Benefits for Women with Breast Cancer. *Journal of Computer-Mediated Communication* 10, JCMC1047 (Jul 2005), 1 pages. https://doi.org/10.1111/j.1083-6101.2005.tb00268.x

[43] Maria Y. Rodriguez and Heather Storer. 2020. A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data*. *Journal of Technology in Human Services* 38, 1 (2020), 54–86. https://doi.org/10.1080/15228835.2019.1616350

[44] Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. 2022. Unpacking Invisible Work Practices, Constraints, and Latent Power Relationships in Child Welfare through Casenote Analysis. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3491102.3517742

[45] Teddy Seyed, Alaa Azazi, Edwin Chan, Yuxi Wang, and Frank Maurer. 2015. SoD-Toolkit: A Toolkit for Interactively Prototyping and Developing Multi-Sensor, Multi-Device Environments. In *Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces (ITS '15)*. Association for Computing Machinery, New York, NY, USA, 171–180. https://doi.org/10.1145/2817721.2817750

[46] Poorna Talkad Sukumar, Ignacio Avellino, Christian Remy, Michael A. DeVito, Tawanna R. Dillahunt, Joanna McGrenere, and Max L. Wilson. 2020. Transparency in Qualitative Research: Increasing Fairness in the CHI Review Process. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381066

[47] Rivka Tuval-Mashiach. 20161027. Raising the curtain: The importance of transparency in qualitative research. *Qualitative Psychology* 4, 2 (20161027), 126. https://doi.org/10.1037/qup0000062

[48] Jennifer Wortman Vaughan and Hanna Wallach. 2021. *A Human-Centered Agenda for Intelligible Machine Learning*. The MIT Press, Cambridge, Massachusetts, Chapter 8, 19. https://doi.org/10.7551/mitpress/12186.003.0014

[49] Ryen W. White and Resa A. Roth. 2009. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1, 1 (Jan 2009), 1–98. https://doi.org/10.2200/S00174ED1V01Y200901ICR003

[50] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (2013-05-13) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 1445–1456. https://doi.org/10.1145/2488388.2488514

[51] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through Design as a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704

Received May 2022; revised August 2022; accepted September 2022