Search with Discretion: Value Sensitive Design of Training Data for Information Retrieval

Modassir Iqbal

College of Information Studies, University of Maryland, College Park, College Park, Maryland, miqbal@terpmail.umd.edu

Katie Shilton

College of Information Studies, University of Maryland, College Park, College Park, Maryland, kshilton@umd.edu

Mahmoud F. Sayed

Computer Science and UMIACS, University of Maryland, College Park, College Park, Maryland, mfayoub@cs.umd.edu

Douglas W. Oard

College of Information Studies and UMIACS, University of Maryland, College Park, College Park, Maryland, oard@umd.edu

Jonah L. Rivera

College of Information Studies, University of Maryland, College Park, College Park, Maryland, jriver15@terpmail.umd.edu

William Cox

College of Information Studies, University of Maryland, College Park, College Park, Maryland, wcox@terpmail.umd.edu

ABSTRACT

This paper describes and assesses the value sensitive design (VSD) of a test collection: data used to train and evaluate a machine learning system for information retrieval. The project used the VSD framework and methods to design a test collection annotated for *discretion*. We conducted qualitative stakeholder interviews to develop values personas, which guided annotation of a collection of corporate emails for contextual notions of sensitivity. Both qualitative and quantitative evaluations of the method reveal that the values personas concretely shaped annotators' sensitivity judgments, and analysis of the test collection itself demonstrates that the sensitivity annotations have utility for identifying features that may correlate with email sensitivity. Values personas for training data annotation expand the toolkit of methods for value-sensitive machine learning.

CCS CONCEPTS

• Information systems → Information Retrieval; • Security and Privacy → Information Retrieval; • Human-centered computing → Human-computer Interaction → HCI theory, concepts and models; • Human-centered computing → Collaborative and social computing

KEYWORDS

Value-sensitive design, training data, machine learning, personas, privacy

1. INTRODUCTION

Curating training data is a critical practice in machine learning. In particular, training and evaluation data can impact the values and impacts of a resulting system [2, 5, 8, 13]. Training data teaches an algorithm which data features are important for decision-making, and evaluation data determines how well a machine learns to perform a meaningful task. But the definitions of "important" and "meaningful" are constructions, reflecting human needs and priorities. Designing human needs and priorities into machine learning is a nascent research area of value-sensitive algorithm design [4, 25, 27], and the best methods for doing so are an open area of research.

This paper expands on existing work in value-sensitive algorithm design by demonstrating and evaluating value-sensitive design (VSD) techniques to shape a test collection: the data used to train and evaluate an information retrieval system. Information retrieval (IR) systems traditionally value relevance (finding and providing what a searcher is looking for) and speed (returning results quickly). But IR systems might also prioritize other values. For example, for sensitive documents or personal data, a retrieval system might value discretion: providing access to some stored information without revealing other, private information.

An IR system that prioritizes discretion has important applications in a world of granular personal documentation. Take the case of providing public access to email records. Email constitutes a plentiful and detailed record of personal and organizational lives. However, there is frequently great reluctance on the part of individuals and organizations to make email available for research. According to a Council of Library and Information Resources report, few archives systematically collect email because of numerous challenges, including sensitivity and donor trust [6]. Public releases of high-profile email collections (such as those of Hillary Clinton and Jeb Bush) have demonstrated significant challenges for providing access while also protecting private information ranging from personal information about correspondents to state secrets.

To address this challenge, our research team is designing a machine learning system to provide search access to email collections while also respecting the contextual and individual privacy preferences of email creators. We refer to the balance between access and protection as *discretion*. To build machine learning systems that support discretion, we need to train classifiers to detect and protect diverse notions of sensitive content. Protecting sensitive content is a huge challenge because notions of sensitivity vary both according to personal preference [18, 21] and social context [15, 20]. Indeed, Mulligan et al. [20] have called privacy "an essentially contested concept": a phenomenon for which disputes about its meaning are a central part of the definition.

This paper explores and evaluates one method of building a test collection to train a search system that responds to a fuzzy, socially dependent value. Value-sensitive design (VSD), a framework developed within Human-Computer Interaction (HCI) to incorporate human values as key design elements, has been applied to a variety of technologies, but applying VSD to the design of test collections (and more broadly, to the development of training data for machine learning) is a relatively new practice [2, 4, 27]. We hope to extend the conceptual and methodological toolkit for value-sensitive algorithm design by asking the following research questions:

1) How can VSD methods be adapted to support curation of test collection training and evaluation data?
2) How does pairing VSD methods of stakeholder analysis with HCI methods of persona creation influence annotator evaluations of sensitivity in training data?

We answer the first question by adapting qualitative interviews and values persona creation techniques to support annotation and create a test collection that usefully represents discretion. Note that our focus here is on a representation that is useful for supporting the development of new retrieval technologies. As George Box reminds us, models are necessarily incomplete representations of the world, so all models are wrong; but some models are useful [3]. We answer the second question by evaluating the resulting test collection. Both qualitative and quantitative evaluations of the annotated data reveal that the values personas concretely shaped annotators' sensitivity judgments, and analysis of the test collection itself demonstrates that the sensitivity annotations have utility for

answering questions about patterns associated with email sensitivity. But the use of the values personas also highlights challenges of context and bias in test collection development.
The paper begins with background on applications of VSD in computer-supported cooperative work broadly and machine learning more narrowly. It also reviews the challenges of email search and sensitivity to justify information retrieval as a site for value-sensitive training data development. The paper next describes the conceptual and empirical investigations that supported our value-sensitive design process. We then describe the primary technical investigation: the use of personas to distill qualitative interview data to support human annotation. We use both qualitative and quantitative analysis to evaluate how the use of two values personas impacted the annotation process, describe the utility of the resulting test collection, and discuss the implications for using values personas to create test collections.

2. BACKGROUND

Friedman et al.'s value-sensitive design (VSD) approach [9, 10, 12] provides a scaffolding for researchers seeking to conceptualize and operationalize complex human values for design. VSD has evolved over time, but recent applications of VSD [7] involve four core commitments: a proactive stance in purposefully designing for values; an acknowledgement that values in design and use co-constitute each other; attention to both direct and indirect stakeholders in design; and a tripartite methodology that uses iterative theoretical, empirical, and technical approaches.
Recent research increasingly grapples with values-sensitive approaches for designing machine learning systems. Muller et al inspire our work by pointing out the conceptual similarities of qualitative data collection (particularly grounded theory) and machine learning, and suggesting research methods that take advantage of both traditions [19]. Holstein et al. [13] point out that ML tools are diverse and varied, and that context-dependent tools and methods are needed to design for human values such as fairness. Yang et al. [26] describes the challenges of user-centered design for AI, reporting challenges in prototyping and iterative testing, and challenges of crafting user experiences around sometimes unpredictable outputs. Chen and Zhu [4] demonstrate that values-sensitive approaches can be used to expand machine learning design values beyond fairness, accountability and transparency, considering a wide variety of values in the domain of learning analytics. They used value-sensitive design stages such as stakeholder analyses, prototyping, and stakeholder evaluations to design recommendation algorithms for crowd work on Wikipedia [4]. Zhu et al. [27] argue that a value-centered method that balances stakeholder needs with automated learning is particularly useful for challenging social problems for which there is no ground truth. Their interventions focus on the algorithms themselves, deciding what they will do and what factors they will consider, and they do not explicitly attempt to shape training data. In the domain of natural language processing, Bender and Friedman use VSD to document biases in natural language training data [2]. Similar documentation procedures for other ML datasets have been suggested by Holstein et al [13].
Steps like identifying stakeholder values, documenting biases in training data, and prototyping value-sensitive algorithmic responses leave ample room for methodological innovation. Value sensitive design has a rich body of work to apply to the challenge of identifying stakeholder values [11], but less about how to respond to those values during machine learning system design. In particular, Zhu et al. cite the challenge of "how to address the fundamental mismatch between human styles of interpretation, reasoning, and inputs and statistical optimizations of high-dimensional data" [27]. While their approach focused on building human values (and humans themselves) into the decision-making loop, we take a different approach: using human values to shape the training data set from which the system learns. This complementary approach, combined with humans in the loop at later stages, will strengthen the values-sensitivity of machine learning systems.
We have chosen to apply VSD to automated systems for email retrieval because of the complex nature of email sensitivity. Since the widespread adoption of email by organizations in the 1990s, email has become a critical tool for business and an important organizational record. However, individuals often

treat email accounts as if they are private, leading to conflicts between workplaces and employees over how much and how closely to monitor work email accounts [14]. Though US legal standards largely dictate that employees have no expectation of privacy in work email, some employers do not wish to erode trust by subjecting workers to undue surveillance. Employers may for example wish to scan email for mission-critical information without accessing personal correspondence [14].
Similarly, email can be a valuable historical record, but only if individuals can be persuaded to let their emails be preserved for research. And few archives have developed programs to preserve email. Writing about "the untapped potential of email archives," a Council of Library and Information Resources report lists privacy and security concerns as among the challenges archivists must face when seeking to collect and preserve email collections [6].
A system that provides access to non-sensitive information while protecting sensitive information could help mitigate both workplace and archival challenges in email search. VSD is particularly appropriate for a search system that values *discretion*, which we define as revealing only contextually appropriate information. Our definition relies on a theoretical framework for defining private or sensitive information based on the notion of *contextual integrity* [22]. Although individuals have different ideas of what constitutes private information, and different levels of sensitivity about privacy, theoretical and empirical work has shown that individuals' privacy concerns are largely shaped by social norms within particular information contexts [17, 21]. Social norms dictate what information it is acceptable to collect, who can have access to it, whether it should be kept confidential, and how it can be shared and reused. In this paper, we use the value *discretion*—access to contextually appropriate information—to describe our overall goal for the search system. We operationalize that value using human judgments about information *sensitivity* in context. The theoretical framework of privacy as contextual integrity supported a methodological approach of first having stakeholders describe and discuss sensitivity in different contexts, and then annotating data in a test collection to reflect those sensitivities. Our approach is described in detail in the next section.

3. STAKEHOLDER ANALYSIS AND VALUES PERSONA CREATION

We began our approach to the value-sensitive design of a test collection using an empirical technique central to the human-centered design tradition: stakeholder interviews. Value-sensitive design asks researchers to consider values of both direct and indirect stakeholders [11]. This paper focuses on a primary stakeholder concerned with email sensitivity: email authors. A primary challenge for our project was narrowing *all email authors* into a set of stakeholders to guide our design process. Because sensitivity concerns vary cross-culturally as well as by context, interviews could be approached in two ways: to maximize diversity, or to target a particular culture and context. We chose the second approach, because focusing on the sensitivity concerns of a similar group of people would reduce the categories and situations considered sensitive, and therefore increase annotators' certainty of sensitivity judgments. Future work can then expand on the categories regarding sensitivity we found through methods better able to capture diversity, such as surveys.[1]
We chose to focus our stakeholder interviews on late-career professionals most likely to be thinking about donating their emails to archives. We first contacted archivists who had processed email collections to find existing donors, but this process confirmed that few archives collect emails to date. As a result, working with archives yielded only four donor interviews, primarily with men (three academics and one industry researcher), all of whom worked in computer science. The research team met to discuss strategies for identifying additional interview subjects, with an eye towards expanding

1 Future work might also improve our design process by considering the values of other direct and indirect stakeholders, such as email recipients, historians, archivists, and managers.

the professional backgrounds in the subject pool. After a conversation with colleagues in our University Archives, we solicited individuals who had been named Distinguished University Professors. This pool of potential subjects was chosen because awardees are likely to be targeted as donors by the University Archives for their "Papers of Faculty and Administrators" collection. We pursued this outreach strategy until we reached conceptual saturation around *discretion concerns* and *discretion practices*, two major themes that grew out of our focus on the value of discretion (described in more detail below). The final pool of participants consisted of five men and four women, all Americans prominent in their careers, and all over the age of 50. This tightly focused sample achieved our goals – conceptual saturation around a focused set of sensitivity concerns – but it also had several limitations, which we discuss below. We have used pseudonyms for participants, and the research was approved by our IRB.

We used Nissenbaum's [22] framework of contextual integrity to form interview prompts that guided individuals through ways to define and identify sensitive information, as well as their broader concerns about discretion in future public or researcher access to their email collections. We recorded all interviews (ranging from 20-60 minutes) and had them professionally transcribed. We used thematic analysis to analyze the rich qualitative data about stakeholders' discretion concerns and the other values they expressed during interviews (such as the values underlying their motivations for donating emails). Two authors coded the interview transcripts using an open coding scheme shaped by our literature review and theoretical framework. Each author coded a matching subset of interviews and then met to discuss themes, questions, and disagreements. After refining our codes, we then re-coded the entire set of interviews using our final coding scheme. Our final coding scheme focused on four areas: the values underlying donors' motivations for considering email donation; their discretion concerns; any actions they had taken to address their concerns; and any potential solutions that they proposed to address email sensitivity challenges.

After coding each interview, we analyzed the distribution of codes for each respondent. All respondents were extensive users of email and used it for a myriad of purposes, but participants differed on two key attributes: their *discretion practices*, or their level of caution while using email (ranging from careful email practices to uninhibited email practices), and their *discretion concerns*, the number, and variety of 'pain points', contexts, or types of sensitive information they believed to be contained in their email (ranging from few to many). We organized these key attributes into four quadrants, described in more detail below.

3.1 Discretion Practices

Interviewed stakeholders described widely divergent email practices. For example, social scientist Nicholas F. discussed his less-careful practices this way:

> *There are just too many things that I did not filter during my use of email that I feel are not appropriate to be available for public discovery... Filtering means that I've said or I've written certain things without, although I should have, considering whether or not those things would be read by people other than the people for whom they were intended.*

Computer scientist Brandon S. described a similar practice:

> *For a number of years, oh, many years, I completely commingled my work emails and my private emails. I didn't start separating them until the late 1990s and as a result, I've got a huge mess on my hands in terms of moving things apart.*

This contrasted with Professor Mary C., who responded:

> *I must say, I never put anything in an email that I would not want on the front page of The New York Times, and that includes anything to do with administrative or opinions of my [chuckle] fellow colleagues, or anything like that. I am very, very, very careful*

> *what I say in emails, and always have been. In terms of, "Would it be embarrassing to anybody to read these?" I think the answer is no.*

A third perspective was provided by Professor Candice C., who had no concern for her own email practices but was concerned about those of others. Early in the interview, she described her email use this way:

> *I go to great lengths not to put anything in an email that I wouldn't see as part of the public record. ...Well, that's the key point, you see. Email is not private. So, it's really quite different from earlier communications. ... So, I don't need any reminding. I might have got a little bit more careful over time, but for basically as long as I remember, yeah. ... One must assume that it's not private.*

But later, she started musing about the privacy of her correspondents:

> *One can imagine a situation where one has a graduate student complaining bitterly about some other person in the field. Yeah, that makes it a little trickier... I don't think I could open my email records.*

3.2 Discretion Concerns

A second central theme in our interviews was the different types of information that the respondents considered sensitive, as well as other contextual elements such as role, transmission principles, and information uses, which influence their concerns about discretion. The respondents spoke about a variety of information in their emails that they considered private. Gossip and personally identifiable information were the most frequently mentioned types of sensitive information. For example, as scientist Wesley P. described:

> *Okay, so probably the thing that would most concern me would be my opinions about individuals. Typically, these would be both scientific and personal opinions. So, I might say of somebody in an email, "By the way, that guy is a total jerk, despite the fact that he is a genius." [chuckle]*

Personally, identifiable information (PII) in the context of research could include identifying information for research participants. As social scientist Nicholas F. described:

> *Although we are extraordinarily careful not to use names or any kind of identifying information in emails when we communicate about that stuff, sometimes, it always slips up. Somebody slips up, and then I get an email about a particular [research] subject or a particular participant, and it has all this confidential information in it.*

Gossip and PII were followed in the sheer number of mentions by information restricted in particular contexts, such as student records in educational settings, trade secrets in corporate and research settings, and human resources information in all workplaces. For example, engineer John R. described sensitive content in his email as a result of working in areas overlapping human resources:

> *So, everything is in there about personnel evaluations, and salary information, and sexual harassment lawsuits, ... gender reassignment issues, and suits from employees against me for dismissing them, and just all kinds of stuff. So, I would have to go through and find all of that.*

Similarly, Mary C. describes the problem of student records in email:

> *I certainly use [email] to communicate with students, and of course, the university sets up course mail distribution lists. I have always communicated with students that way. I send students assignments; they send me papers with their open book exams and so forth. That all comes to me via email.*

A common concern among multiple respondents was the intermixing of their work emails and personal, social, or private conversations (often coded as "intimacies," "emotional content," and "health and financial information"). As computer scientist Brandon S. described:

> *... things that I would consider personal, that is my business and not anybody else's business. ... I'm aware that journalists and others and historians like these juicy personal stories….. My relationships with my sister, my daughters, others, would have been further concerns.*

Another major concern was the roles of people who access their emails in the future and their intent of use. As computer industry researcher Daniel L. put it:

> *I think the concern I would have would be with somebody who is using it to invade someone's privacy.*

Participants felt that journalists, investigators, or lawyers might try to find information to either harm or mislead someone or build up a story that could easily have been misinterpreted due to their lack of understanding of the context. Computer scientist Danielle E. raised the concern of loss of context when emails are viewed by outsiders. She described:

> *Yeah, I do think that it's really important to represent the communication appropriately... You know, some [emails] go to large numbers of people, and some go to particular others in a really private communication. So, what somebody communicated privately versus what they said aloud, I would want treatments of that to be included.*

Worry about harm to others also appears as a consistent theme, although the roles perceived to be affected are disparate. As scientist Thomas J. put it:

> *I imagine there are people who, either intentionally or totally inadvertently, have information in their email about other people that could harm the other people if it became public.*

Industry researcher Daniel L. was a bit more specific about who might be at risk in his email:

> *Yeah, I think my primary concern would be, somebody, invading the privacy of my family or friends. Again, once I'm dead, I don't care but somebody who might... A stalker, ex-boyfriend or girlfriend or a friend, or a step kid or something like that. So that's the kind of thing, somebody who would be trying to get information, that they shouldn't have about somebody I care about.*

Finally, our respondents almost universally expressed a concern regarding the time and effort it would require to filter through their email collections to prepare them for donation. As scientist Wesley P. put it:

> *...there's no way I'm ever ever going to go through my emails and decide which ones are personal and which ones aren't as supposedly Hillary Clinton and Ivanka Trump have done. And that's... I'm just never going to do it because I don't have time.*

We combined worries about information types, roles, user intentions, and time and effort to create the code *discretion concerns*, which reflect the diversity of information factors that stakeholders identify as contributing to sensitivity in email collections.

3.3 Values Persona Creation

With discretion practices and concerns identified as key themes, we were able to group our interview subjects into one of four quadrants as shown in Table 1.

Table 1. Values Persona Types Based on Discretion Practices and Concerns

	Many Discretion Concerns	Few Discretion Concerns
Uninhibited Practices	Diarist	Unconcerned
Careful Practices	Cautious	Protected

For example, the *diarist* was uninhibited in their email practice and as a result, felt their email contained many types of sensitive information or could easily be used by others in indiscrete ways. In contrast, the *unconcerned emailer* was similarly uninhibited in their email practice but did not worry about having many kinds of sensitive information in their email or potentially indiscrete uses of their email, perhaps because they felt they had little to hide. The *cautious emailer* felt they had been circumspect in their own email communication, but still discussed a large variety of sensitive information that may have ended up in their email collections despite their practices. (They usually recognized that others whom they corresponded with had been less circumspect). And finally, the *protected emailer* had similarly cautious practices, but as a result, felt their email to be relatively free of sensitive information and had relatively few concerns about how others might use that email.

Interestingly, we did not have anyone in our interview sample who fell into the 'unconcerned' quadrant. Indeed, it seems unlikely that there are few people with distinguished careers who have few concerns about sensitive email *and* have been uninhibited in their email practices. And while we did have respondents who fell into the 'protected' quadrant, we chose to initially focus *values persona* creation on the two quadrants (the diarist and the cautious emailer) with the largest variety of discretion concerns to support a robust variety of 'sensitive' classifications in our annotations. Personas are a realistic description of an archetypal representation of a system's users, used to synthesize complex qualitative data for design. Personas typically represent user goals and attitudes [1] and in VSD, they can be developed using frameworks that highlight and derive user values from qualitative research and participatory-design [16]. Our personas distilled composite insights from our interview subjects to help human annotators make judgments about how real-world email senders might classify sensitive information (Table 2 and 3). We added additional detail to the values personas drawn from the composite profiles from our interviews, including the professional background of the interview subjects, their imagined audiences for their emails, their motivations for email donation, and potential solutions for navigating email sensitivity. This process led to the creation of two values personas to support value-sensitive training data annotation: Holly Palmer (the diarist emailer) and John Snibert (the cautious emailer).

Table 2. Values Persona Discretion Practices

Name	Discretion Practices
Holly Palmer	Holly has used her work email to communicate with both personal and professional connections about work, home, logistics, gossip, and trade secrets.

John Snibert	John has used email extensively for his work, including coordinating projects, planning presentations, and having conversations with his company and other business leaders. He has also used email for sensitive matters like conversing with family and romantic partners. He believes he has been careful about what he puts in his email, and he has already done some curating and deleting of sensitive information.

Table 3. Values Persona Discretion Concerns

Name	Discretion Concerns
Holly Palmer	She is worried about the time and effort it will take to filter and delete conversations about her family and personal life. She also mentions travel and other financial receipts, and professional reviews of colleagues as information that she would not want to be shared. She is worried about potential harm to her and others' reputations (family and colleagues) because of the unflattering things she has written about them in her emails in the past. She is worried that her emails will be taken out of context.
John Snibert	John is aware that there are sensitive emails in his collection that include his conversations with his family and romantic partners, opinions about his peers, and collaboration on projects that contain proprietary information and trade secrets. He worries about the intentions of the people who might access his emails, like journalists looking for a story.

4. ANNOTATING A TEST COLLECTION

After we developed the values personas to reflect how our stakeholders operationalized *discretion* through their concerns and practices, we used the values personas to guide a pair of annotators as they built a *test collection*. The terminology comes from information retrieval, a discipline traditionally focused on finding relevant documents in response to a user's query. Test collections are used for laboratory studies that typically precede the use of live systems with real users, and the usual focus of their use is on the effectiveness (rather than the efficiency) of the retrieval technique. Test collections

can be used either as training data or as test data; or, if split, part of the collection can be used for each purpose. Typical test collections are built by first choosing a corpus of documents (in this case, emails). Next, the team creates topics, which serve as examples of the kinds of information searchers might seek. In typical practice, these topics contain three fields: 1) *title*, a short Web-like query; 2) *description*, a natural-language sentence that a searcher might use to ask for the same information; and 3) *narrative*, guidelines for judging relevance. Human annotators then make judgments about the degree to which particular documents are relevant to a given topic. After judging relevance guided by topics, each annotator made two sensitivity judgments: one based on their interpretation of the values of their assigned values persona, and a second based on their *personal* interpretation of sensitivity. We describe each step in our process, as well as how those steps were guided by VSD considerations, in more detail below.

4.1 Finding a corpus

An email test collection requires an existing set of real-world emails available for research. Our personas were generated by interviews with academics, and we had originally planned to build a test collection with the email corpus of a distinguished professor to which we have been granted research access. However, reflecting on our interest in discretion revealed that this plan would require outside annotators (in our case, undergraduate researchers) to view and label sensitive content from a faculty member. Deeming this risk unacceptable to the ethos of our project, we turned to a dataset purpose-built with protections for research use: the Avocado Research Email Collection, the contents of late-1990s email accounts from a now-defunct information technology company, distributed under a restricted license for research by the Linguistic Data Consortium [23]. However, the emails in the Avocado collection, from a business, rather than academic, context, had a contextual mismatch with our academic personas. We discuss this limitation further in our evaluation section.

4.2 Topic creation

The information retrieval researchers on our team guided the choice of the topics for the Avocado emails, as the returned results would need to be useful for judging both relevance and sensitivity. With this constraint in mind, three undergraduate students created 137 topics designed to be relevant to many parts of an individual's life, such as business-related matters (e.g. promotions), events current to the time period (e.g., the Olympics, or the massacre at Columbine High School), personal matters (e.g., drug use, or vacations), and topics might invoke sensitivities (e.g., keywords such as selfish or tired) in a business email setting. The annotators created title, description and narrative fields for each topic. We next formed pools of emails for annotation by combining the potentially relevant documents identified during topic creation with additional search results from 18 automatic search systems, described in more detail in [blinded for review]. The resulting pools have about 100 emails, on average, per topic.

4.3 Making judgments

We employed two of the three topic creators as annotators.[2] Annotators used these topics to query the email collection and examine the query results to see if the emails returned were 1) relevant to the topic; 2) sensitive according to assigned personas, and 3) sensitive according to their own judgment. The Avocado collection has emails from many creators; all were annotated as if they were created by the persona. We displayed the values persona descriptions in a user interface that two annotators used to label the resulting pools of emails. Within the values persona, the persona's use of email

2 Both annotators later also contributed to analysis of inter-annotator agreement after annotation was complete, and have thus joined us as authors on this paper.

(discretion practices) and pain points (discretion concerns) were highlighted for added visibility. [1–4]. shows one sample of the annotation task and values persona. We collected the personal judgment of the annotator regarding sensitivity as a point of comparison to help us determine whether the values personas created to represent late-career, long-term email users would produce different judgments than those made intuitively by early-career annotators (undergraduate students).

Our annotation process sought to balance reliability (measured through inter-annotator agreement) and scale (measured as the total number of emails annotated). We began annotation with a four-round training period. In each round, the two annotators were given the same values persona, and the same set of emails retrieved from two search topics. Annotators ran three rounds using the John Snibert values persona, and held meetings in between rounds to address questions of interpretation (e.g., the annotators agreed that when presented with a case for which no useful guidance was available from the values persona description, they would interpret sensitivity using their understanding of common societal expectations regarding sensitivity). Both annotators achieved substantial agreement on relevance and sensitivity judgments (0.76 kappa on relevance, 0.66 on sensitivity). The Kappa coefficient measures chance-corrected inter-annotator agreement. The range of possible values of kappa is from -1 to 1. When Kappa is negative, it indicates disagreement. When Kappa is positive, Landis and Koch [15] suggest that it can be interpreted as follows: 0.00–0.20 as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement, 0.61–0.80 as substantial agreement, and 0.81–1.00 as almost perfect agreement. These results indicate annotators arrived at substantial agreement on material that John Snibert would have felt was sensitive, indicating that his values persona was well defined, and both annotators understood it similarly. Finally, we ran one round of sensitivity annotation training using the Holly Palmer values persona, which achieved Kappa of 0.53 (moderate agreement) on sensitivity. After discussing their differences, we expected annotators to have even higher agreement, so from that point forward one of the annotators continued to annotate as John Snibert, and the other as Holly Palmer. The final test collection has 60 topics, divided evenly across the two personas, and an additional five topics shared between personas. Each persona was used for 35 topics, and each topic is associated with a set of emails judged for relevance to that topic as well as sensitivity in the opinion of the specified values persona.

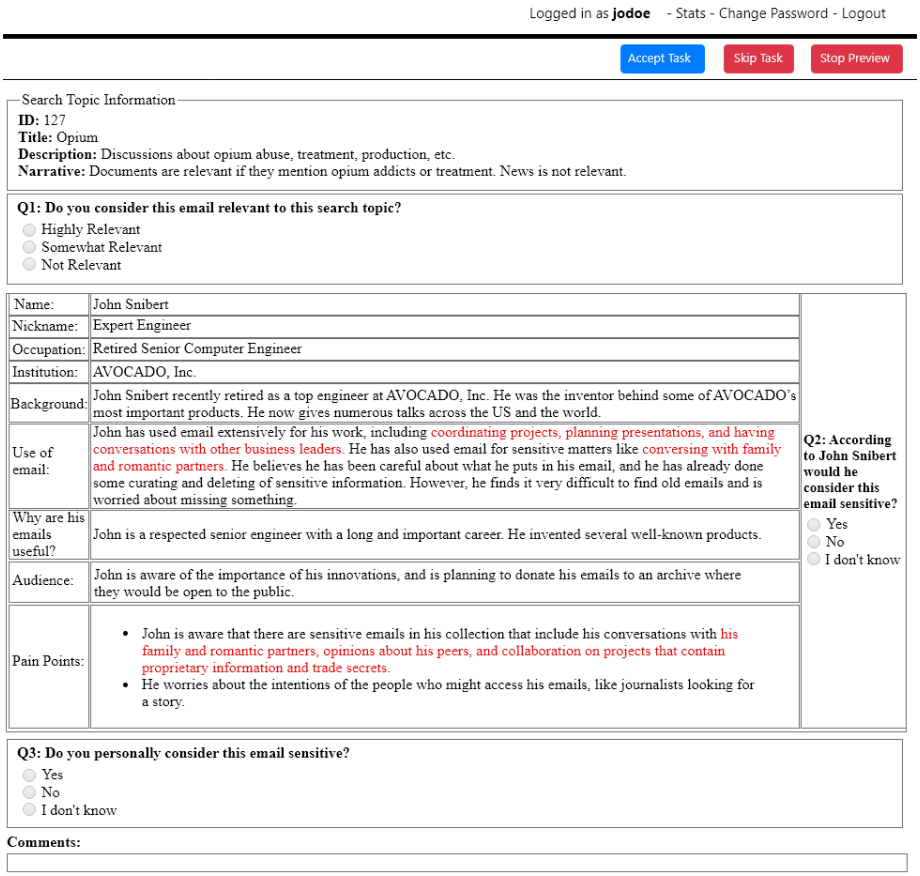

Fig. 1. The email annotation interface (synthetic message and attachment created for public dissemination)

5 EVALUATION

We evaluated the utility and effect of the values personas in three ways: 1) through reflective interviews with the annotators; 2) by computing Kappa coefficients to determine the differences between the annotators' personal judgments of sensitivity and their judgments from the perspective of the values personas, and 3) by evaluating the utility of the test collection built with the sensitivity annotations. We discuss each of these evaluations in more detail below.

5.1 Reflective interviews

Interviews revealed that the values personas provided a helpful rubric to assist annotators with sensitivity decision-making. Annotators reported that they read each values persona description multiple times as they evaluated the sensitivity of any given email, sometimes reading it again after making an initial decision to validate the assessment. They frequently used the values persona's discretion concerns, discussion of their email practices, and background sections of the values persona description to assess content sensitivity.

5.2 Computing differences between sensitivity judgments

The quantitative analysis of judgments supported the qualitative findings, as the annotators tended to mark more content as sensitive when using a values persona than when annotating for their own

opinion of sensitivity, particularly when there were many sensitive documents for a topic. Tables 4 and 5 illustrate the raw count differences and the Kappa coefficients between the personal and the values persona sensitivity judgments. For most topics, the number of documents coded as sensitive was higher when using the values personas than when relying on the annotator's personal opinion of sensitivity (a handful of exceptions were topics in which very few documents were sensitive). Kappa values were generally low, indicating only slight agreement between the annotator and the values persona. Although there are some topics in which the annotator and values persona had significant agreement on sensitivity, many of these topics had very few documents, so even a few annotator-persona agreements could result in a high kappa coefficient. However, a few topics (e.g., 'terrorism' for John's values persona, or corruption for Holly Palmer) have higher agreement despite having more documents and may represent topics on which sensitivity concerns were seen as more obvious or universal by the annotator.

Table 4. Percentage and number of sensitive documents for John Snibert persona and annotator

Topic Title	Sensitivity agreement (Kappa)	Percent of sensitive emails (persona)	Number of sensitive emails (persona)	Percent of sensitive emails (annotator)	Number of sensitive emails (annotator)	Number of annotated emails
Depreciation	-0.07	58%	51	3%	3	87
Viagra	0.00	100%	33	6%	2	33
Parents	+0.04	93%	154	23%	39	164
Storage Space	+0.05	48%	40	2%	2	82
Enron	+0.06	86%	46	16%	9	53
Fax	+0.09	49%	75	5%	9	151
Mortgage	+0.11	41%	36	8%	7	87
Movies	+0.11	40%	62	3%	6	155
Presidential	+0.13	15%	12	1%	1	79
N word without r	+0.14	67%	106	13%	21	157
Communism	+0.15	26%	44	3%	6	168
Electricity	+0.16	3%	4	5%	6	108

Cats	+0.16	62%	55	12%	11	88
Meetings	+0.16	81%	145	26%	48	179
Respect	+0.17	66%	90	15%	21	136
Lord of the rings	+0.18	13%	9	1%	1	69
Chechnya	+0.18	11%	9	1%	1	79
Vacation	+0.19	52%	57	31%	34	109
Bill Gates	+0.19	24%	15	3%	2	62
Promotion, raise	+0.20	8%	12	11%	17	143
Opium	+0.22	23%	26	3%	4	111
SARS	+0.25	65%	52	21%	17	80
Jury Duty	+0.26	68%	28	24%	10	41
Avocado stock	+0.27	12%	14	4%	5	113
Stealing	+0.35	40%	51	12%	16	126
Google	+0.38	27%	20	8%	6	74
MDMA	+0.40	63%	55	33%	29	86
China	+0.42	21%	16	10%	8	74
Laid-off	+0.50	42%	39	24%	22	91
Sexual Harassment	+0.52	56%	27	31%	15	48
Depression and resources	+0.54	45%	27	23%	14	60
Cancer	+0.56	42%	44	22%	23	104
Apple	+0.60	67%	46	47%	32	68
Loan	+0.61	56%	45	36%	29	79
Terrorism	+0.76	32%	27	22%	19	83

Table 5. Percentage and number of sensitive documents for Holly Palmer persona and annotator

Topic Title	Sensitivity agreement (Kappa)	Percent of sensitive emails (persona)	Num. of sensitive emails (persona)	Percent of sensitive emails (annotator)	Number of sensitive emails (annotator)	Number of annotated emails
Impeachment	-0.04	2%	2	7%	6	83
Immigration	0.00	7%	4	0%	0	51
Columbine	0.00	2%	1	0%	0	42
Debt	0.00	0%	0	0%	0	82
Scam	0.00	0%	0	0%	0	35
Storage space	0.00	0%	0	0%	0	80
CPU	0.00	0%	0	0%	0	134
Abortion	0.00	0%	0	0%	0	75
Baby	+0.02	85%	93	4%	5	109
Vacation	+0.03	42%	46	0%	1	109
San Francisco Giants	+0.10	20%	15	1%	1	72
Superbowl	+0.16	31%	24	3%	3	77
Illness	+0.24	29%	27	5%	5	92
Fortune 500	+0.26	11%	12	1%	2	104
Prison/jail	+0.42	10%	7	2%	2	67
Assets	+0.44	13%	13	4%	4	98
Office Parties	+0.48	8%	9	2%	3	111

Cigarettes	+0.49	12%	10	6%	5	80
Napster	+0.53	15%	10	6%	4	66
Yahoo	+0.54	15%	22	6%	9	140
Oakland Raiders	+0.57	16%	9	7%	4	55
Personal	+0.68	16%	29	9%	18	181
Favor	+0.72	6%	11	4%	8	164
Dot-com bubble	+0.73	9%	5	5%	3	54
Tired	+0.74	25%	35	16%	23	138
Parents	+0.76	53%	89	41%	69	167
Profit	+0.83	5%	7	3%	5	140
Accuse	+0.93	18%	18	18%	18	98
Fax	+0.94	6%	10	5%	9	152
Bankruptcy	+1.00	3%	6	3%	6	170
Chechnya	+1.00	4%	4	4%	4	81
Corruption	+1.00	20%	18	20%	18	90
Smartphones	+1.00	3%	4	3%	4	127
Capitalism	+1.00	1%	3	1%	3	180
Republican	+1.00	54%	32	54%	32	59

The annotators acknowledged that they were conservative when using the values personas. As one of the annotators described:

> *I believe I was cautious and tended to mark things sensitive if I wasn't 100% sure about the sensitivity of the email. If there was something even slightly related to the pain points of my persona, I would mark the email sensitive.*

Table 6 shows that agreement between the two values personas (as annotated by their respective annotators) was expectedly low on each of the five topics that were annotated for both values personas. The two values personas thus seem to have successfully captured different sensitivities. In general, the John Snibert values persona (the cautious emailer) generated many more judgments of sensitivity. This may have occurred because John's discretion concerns were tailored for business rather than academic environments and were therefore a better fit for the Avocado email collection.

Many of Holly's concerns, such as the sensitive reviews of colleagues common in academia, would not have appeared as frequently in this collection. As an annotator described the challenge:

> *It was strange, however, marking the trade secrets of a computer programmer as private for an economist. It seems to me that what an economist finds to be a trade secret (research/data) is different from what a programmer considers a trade secret (code/algorithms).*

In addition, despite having modeled John Snibert as cautious about what he put in email, the Avocado collection aggregates the practices of many people, and thus does not actually reflect cautious emailing practices and deletion habits. We assume that John's real-world email (if John were real!) would presumably turn up less sensitive information.

Table 6. Common topics across John Snibert and Holly Palmer annotators

Topics	Sensitivity agreement (Kappa)	Percent of emails marked sensitive – John	Percent of emails marked sensitive – Holly
Vacation	-0.17	52%	42%
Storage Space	0.00	48%	0%
Parents	+0.08	93%	53%
Fax	+0.11	50%	6%
Chechnya	+0.34	11%	4%

5.3 Evaluating the utility of the test collection

Our final evaluation of the values personas was determining the utility of the resulting annotations. A preliminary analysis of sensitivity annotations in the test collection reveals a variety of structural factors that correlate with annotations of sensitivity. These correlations suggest that the annotations produced through the values persona method are meaningful and interesting for suggesting future features we might use to build classification models to *predict* email sensitivity.

For example, we partitioned the email collection by hour of the day in the sender's time zone (0-1, 1-2, ..., 23-24), and then aggregated all judged emails in each time range. Fig. 2. shows the fraction of emails sent in each period that were sensitive for John Snibert (results for Holly Palmer show a similar pattern). As can be seen, the fraction of emails that are sensitive increases later in the day (e.g., after noon or in the evening). And while the total number of total emails sent on weekends is expectedly much lower than the number sent on weekdays, many of the weekend emails are sensitive. In the reflective interviews, our annotators reported rarely referring to the time of day or day of the week on which an email was sent. Because these emergent results comport well with our intuition, we see this as evidence that the sensitivity annotations can lead to discovery of potentially useful patterns in our test collection.

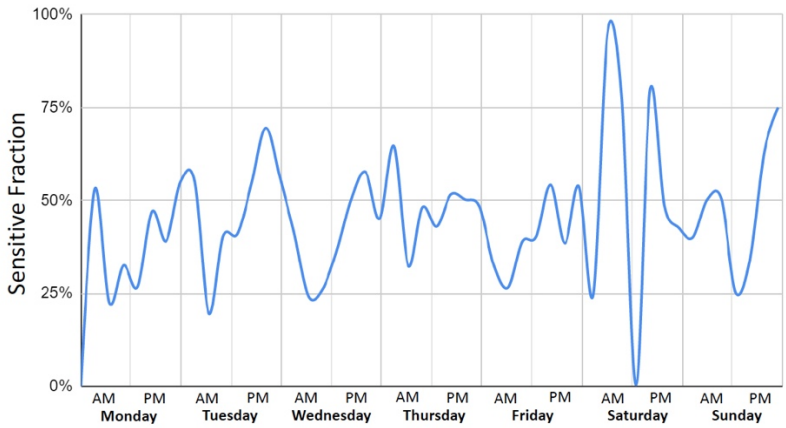

Fig. 2. Density plot of day and time effects on fraction of emails that are sensitive (John Snibert).

Sensitivity also varies based on whether an email sender is inside or outside of the organization. We analyzed the relationship between email address type and sensitivity. We considered an email address internal if the domain of the email address was @avocadoit.com.[3] All other email addresses were considered external. First, we analyzed the address type of email sender. We aggregated emails sent from each address type and computed the percentage of sensitive emails over all judged emails. We then performed a similar analysis for recipients, distinguishing between the case in which all recipients were external and the case in which at least one recipient was internal. As Table 7 shows, emails sent from an internal address were more likely to be sensitive. However, we noticed that some of the emails sent by external addresses were newsletters, which are less often sensitive. The fraction of emails that were sensitive was also higher when at least one of the recipients was inside the company.

Table 7. Effect of email address domain type on fraction of emails that are sensitivity

Name of Persona	Sender		Recipient	
	Internal Domain	External Domain	Some Internal Domains	All Domains External
John Snibert	65.2%	30.0%	53.2%	30.8%
Holly Palmer	30.0%	10.0%	17.9%	14.9%

3 In the Avocado collection, the corporation's actual domain name was replaced by avocadoit in every message.

In both cases, we find that the results of the analyses comport well with intuition and support the conclusion that sensitivity annotations produced using the values personas reflect potentially useful phenomena. The resulting test collection can thus be expected to be useful for its intended purpose: training and evaluation of systems that exercise discretion about which documents should be shown to a searcher [24, 28].

6. DISCUSSION

Creating a collection of emails annotated for sensitivity shows the challenges and possibilities of building values-sensitive test collections. Our first research question asked how VSD methods could be adapted to support test collection training and evaluation data curation. VSD methods influenced numerous training data design decisions. First, VSD influenced our choice of data to use, as we opted for a protected email collection (the Avocado emails) over a more current and arguably relevant but .unprotected data collection (that of a faculty member). Choice of training data is a critical issue for machine learning, and the reflective stages built into VSD can and should shape that choice.

Second, by borrowing personas, a tool from interaction design, to translate stakeholder concerns into concrete annotation decisions, we were able to build a test collection of emails annotated for sensitive content. We demonstrate that the process of building a value-sensitive test collection is one of reduction: moving from complex human preferences to machine-interpretable decisions. Qualitative interviews, thematic coding, and values persona construction offer a method by which to handle that reduction while maintaining some richness and diversity, just as they have in other forms of human-centered design practice. In particular, the richness our personas supported was an expanded range of contextual sensitivities for our annotators to identify and label. While a test collection could have been developed by asking stakeholders to annotate data themselves, creating values personas combined the concerns of multiple individuals, and therefore highlighted a wider range of contextual sensitivities than any one stakeholder might have identified on their own. Our method innovates in an area of machine learning - training data curation - that is emerging as an important area of research.

Our second research question asked how using persona creation would influence annotator evaluations of sensitivity in training data. Our evaluation demonstrates that user-centered design techniques can be applied to training data curation in a way that produces an impact on annotation: values personas helped annotators to produce more judgments of sensitivity. The values-oriented annotations also produced useful results for machine learning development. Using our annotations to look for patterns in the Avocado email dataset revealed that sensitivity correlated with structural features of email, including time sent and whether authors were emailing colleagues or outsiders. These correlations suggest features that may be useful to a machine learning tool.

Finally, our research suggests that value sensitive design of a test collection may be difficult to achieve without qualitative work. VSD for test collections requires interpretation, and that interpretation must be based on the messy notions of human values. For well-established tasks such as annotation for relevance, the tacit knowledge needed for value sensitive design might be gained from hiring expert annotators, as is common in information retrieval and computational linguistics. But when annotating less well characterized phenomena, rendering tacit knowledge explicit through values personas provides an attractive alternative. Machine learning teams interested in value sensitive design for test collections should consider collaborating with experts in social science research.

7. LIMITATIONS

Our choice to facilitate reliable annotation by creating a very focused set of personas from a homogenous interview sample creates the major drawback of our approach and highlights a limitation of qualitative methods for training data curation more broadly: representativeness and diversity. Our

initial plans to develop a system for academics led to interviews conducted in one context, but this complicated training data curation when we decided (for sensitivity reasons) to use data from a different context for annotation. Our annotators reported that they struggled with whether some categories of information should be considered sensitive because they were unique to the industry context. Despite this limitation, however, the results show that many categories of sensitivity translated well between the contexts, and that our annotators were able to use the values personas successfully.

The limited diversity of our personas also intersects with a challenge foundational to value-sensitive machine learning: training data bias. Our interview sample of senior academics means that we have only captured the sensitivity concerns of successful, older, white, American researchers. And our use of the Avocado emails means that the training dataset doesn't represent any information (e.g., secret online accounts, online trolling, burner phones) that may have become sensitive since the late 1990s. Although one advantage of the use of personas in HCI is that they can be diversified to represent viewpoints a design team might not otherwise consider [1,5,7,9] we did not use personas in this way in our research. We did not want to assume the concerns of people not represented in our interviews by creating fictional personas. Collecting empirical data about contextual email sensitivities from a diverse population will require very different methods. A necessary next step will be targeted focus groups or interviews with diverse demographic groups to understand the range of ways discretion concerns might be operationalized by diverse email authors.

We chose to have annotators code for both their own personal ideas of sensitivity, as well as an interpretation of sensitivity based on the values personas. While this afforded the benefit of providing a comparison measure to ensure that coders were using the personas (instead of *only* their own definitions of sensitivity), this also introduced a limitation. There may have been interaction effects between personal and persona judgments: coders may have been influenced in their own judgments by the personas' suggestions of sensitive information. Ultimately, however, analysis shows that personal and persona judgments were significantly different for many topics, despite any interaction effects.

Finally, the process of reducing complex human values into annotations on documents remains necessarily imperfect and should be supplemented with other VSD methods throughout algorithm design.

8. CONCLUSION

Value-sensitive machine learning is a challenging intellectual project that will need to incorporate a rich variety of methods developed over time, much as value-sensitive design has done over three decades. Appropriate methods for the design of training and evaluation data and algorithms will depend on their functions and context, and the values and bias issues likely to be faced in those contexts.

To explore methods for value-sensitive test collection curation we have developed one technique for building the value of *discretion* into a test collection: using qualitative interviews distilled into values personas to guide data annotation. This process was useful for developing a test and evaluation collection of emails labeled for rich, contextual notions of sensitivity, and we demonstrated that using values personas helped annotators label a wider variety of sensitive information than they would have on their own. But the process also built particular biases into that test collection based on the sampling we used to conduct interviews.

ACKNOWLEDGMENTS

We thank our interview participants, whose time and experience greatly shaped this work; Sarah Gilbert and Karen Boyd, who provided feedback on early drafts; and the anonymous reviewers who helped us improve this paper. This research was supported in part by NSF Award IIS-1618695.

REFERENCES

[1]Tamara Adlin and John Pruitt. 2010. *The essential persona lifecycle: your guide to building and using personas*. Morgan Kaufmann, Amsterdam; Boston. Retrieved May 26, 2020 from http://www.books24x7.com/marc.asp?bookid=37227

[2]Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Trans. Assoc. Comput. Linguist.*6, (2018), 587–604. DOI:https://doi.org/10.1162/tacl_a_00041

[3]George E. P. Box. 1976. Science and Statistics. *J. Am. Stat. Assoc.*71, 356 (1976), 791–799. DOI:https://doi.org/10.2307/2286841

[4]Bodong Chen and Haiyi Zhu. 2019. Towards Value-Sensitive Learning Analytics Design. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 343–352. Retrieved from https://doi.org/10.1145/3303772.3303798

[5]Andrew Clement. 1990. Cooperative support for computer work: a social perspective on the empowering of end users. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work - CSCW '90*, ACM Press, Los Angeles, California, United States, 223–236. DOI:https://doi.org/10.1145/99332.99357

[6]Council on Library and Information Resources. 2018. *The Future of Email Archives A Report from the Task Force on Technical Approaches for Email Archives*. Council on Library and Information Resources. Retrieved from https://clir.wordpress.clir.org/wp-content/uploads/sites/6/2018/08/CLIR-pub175.pdf

[7]Janet Davis and Lisa P. Nathan. 2015. Value Sensitive Design: Applications, Adaptations, and Critiques. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, Jeroen van den Hoven, Pieter E. Vermaas and Ibo van de Poel (eds.). Springer Netherlands, Dordrecht, 11–40. DOI:https://doi.org/10.1007/978-94-007-6970-0_3

[8]Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), Association for Computing Machinery, New York, NY, USA, 329–338. DOI:https://doi.org/10.1145/3287560.3287589

[9]Batya Friedman, Alan Borning, Janet L. Davis, Brian T. Gill, Peter H. Kahn, Travis Kriplean, and Peyina Lin. 2008. Laying the Foundations for Public Participation and Value Advocacy: Interaction Design for a Large Scale Urban Simulation. In *Proceedings of the 2008 International Conference on Digital Government Research* (dg.o '08), Digital Government Society of North America, 305–314.

[10]Batya Friedman and David Hendry. 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, ACM Press, Austin, Texas, USA, 1145. DOI:https://doi.org/10.1145/2207676.2208562

[11]Batya Friedman, David G. Hendry, and Alan Borning. 2017. A Survey of Value Sensitive Design Methods. *Found. Trends® Human–Computer Interact.*11, 2 (2017), 63–125. DOI:https://doi.org/10.1561/1100000015

[12]Batya Friedman, Peter Kahn, Alan Borning, Ping Zhang, and Dennis Galletta. 2006. Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics*. DOI:https://doi.org/10.1007/978-94-007-7844-3_4

bib id="bib13" type="Other"><number>[13]Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst. - CHI 19* (2019), 1–16. DOI:https://doi.org/10.1145/3290605.3300830

[14]Meir Hornug. 2005. Think Before You Type: A Look at Email Privacy in the Work Place. *Fordham J. Corp. Financ. Law* 11, 1 (January 2005), 115.

[15]J Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (March 1977), 159. DOI:https://doi.org/10.2307/2529310

[16]Justin Larner. Value-led Personas: A Methodology to Promote Sustainable User-centered Design? 4.

[17]Kirsten Martin and Helen Nissenbaum. 2016. Measuring Privacy: An Empirical Test Using Context to Expose Confounding Variables. *Columbia Sci. Technol. Law Rev.*18, 1 (2017 2016), 176–218.

[18]Kirsten Martin and Katie Shilton. 2016. Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices. *Inf. Soc.*32, 3 (May 2016), 200–216. DOI:https://doi.org/10.1080/01972243.2016.1153012

[19]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), Association for Computing Machinery, New York, NY, USA, 3–8. DOI:https://doi.org/10.1145/2957276.2957280

[20]Deirdre K. Mulligan, Colin Koopman, and Nick Doty. 2016. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*374, 2083 (December 2016), 20160118. DOI:https://doi.org/10.1098/rsta.2016.0118

[21]Alison R. Murphy, Madhu C. Reddy, and Heng Xu. 2014. Privacy practices in collaborative environments: a study of emergency department staff. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (CSCW '14), Association for Computing Machinery, Baltimore, Maryland, USA, 269–282. DOI:https://doi.org/10.1145/2531602.2531643

[22]Helen Nissenbaum. 2009. Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

[23]Douglas W. Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. 2015. Avocado Research Email Collection - Linguistic Data Consortium. Retrieved October 15, 2020 from https://catalog.ldc.upenn.edu/LDC2015T03

[24]Mahmoud F. Sayed and Douglas W. Oard. 2019. Jointly Modeling Relevance and Sensitivity for Search Among Sensitive Content. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, France.

[25] Steven Umbrello. 2019. Beneficial Artificial Intelligence Coordination by means of a Value Sensitive Design Approach. *Big Data Cogn. Comput.*3, 1 (2019), 5.

[26]Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. DOI:https://doi.org/10.1145/3313831.3376301

[27][27]Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.*2, CSCW (November 2018), 1–23. DOI:https://doi.org/10.1145/3274463

[28]Mahmoud F. Sayed, William Cox, Jonah Lynn Rivera, Caitlin Christian-Lamb, Modassir Iqbal, Douglas W. Oard, and Katie Shilton, 2020. A Test Collection for Relevance and Sensitivity. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and*

Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, 1605-1608.
DOI:https://doi.org/10.1145/3397271.3401284