"It's Like the Value System in the Loop": Domain Experts' Values Expectations for NLP Automation

Dilruba Showkat
Northeastern University
Boston, MA, USA
showkat.d@northeastern.edu

Eric P. S. Baumer
Lehigh University
Bethlehem, PA, USA
ericpsb@lehigh.edu

ABSTRACT

The rise of automated text processing systems has led to the development of tools designed for a wide variety of application domains. These technologies are often developed to support non-technical users such as domain experts and are often developed in isolation of the tools primary user. While such developments are exciting, less attention has been paid to domain experts' expectations about the values embedded in these automated systems. As a step toward addressing that gap, we examined values expectations of journalists and legal experts. Both these domains involve extensive text processing and place high importance on values in professional practice. We engaged participants from two non-profit organizations in two separate co-speculation design workshops centered around several speculative automated text processing systems. This study makes three interrelated contributions. First, we provide a detailed investigation of domain experts' values expectations around future NLP systems. Second, the speculative design fiction concepts, which we specifically crafted for these investigative journalists and legal experts, illuminated a series of tensions around the technical implementation details of automation. Third, our findings highlight the utility of design fiction in eliciting not-to-design implications, not only about automated NLP but also about technology more broadly. Overall, our study findings provide groundwork for the inclusion of domain experts values whose expertise lies outside of the field of computing into the design of automated NLP systems.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**; **Computer supported cooperative work**.

KEYWORDS

Design Fiction, Participatory Design Fiction Workshop, Natural Language Processing (NLP) Automation, Values in Automated NLP

ACM Reference Format:
Dilruba Showkat and Eric P. S. Baumer. 2022. *"It's Like the Value System in the Loop"*: Domain Experts' Values Expectations for NLP Automation. In *Designing Interactive Systems Conference (DIS '22), June 13–17, 2022, Virtual Event, Australia.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3532106.3533483

1 INTRODUCTION

In recent years, we observe an increasing growth and interest in the design and development of automated AI/ML [85, 135–137], automated data science (e.g., AutoDS [133]), and language processing systems [17, 35], including large models [e.g., 30, 46]. Some of these language technologies are designed to be task-agnostic [30, 46], meaning they are suitable to perform a wide range of tasks in diverse application domains. Examples include natural language generation of news articles [46], question-answering, query understanding [1], translation and summarization [92], and others. Several of these tools (e.g., AutoML/AutoAI/AutoText) are often developed with dual objectives in mind to facilitate both technical experts and domain experts [35, 104, 135, 137]. For example, with the help of these tools, expert data scientists no longer have to engage in repeated low level programming tasks, while domain experts such as business professionals can build and run ML/NLP models easily [135]. In general, researchers often emphasize reducing the entry barrier in ML/NLP as a rationale justifying these automated technology developments [35, 135].

However, only limited attention has been paid to the values that guide the design of those systems, let alone how those values might play out in implementation. While automated developments enhance human efficiency, potential, and capabilities, they also raise an increasing concern of possible harm, misuse, and cruelty [14, 56, 74, 88, 92, 100, 103]. Neither ML technologies nor the field of ML in general are value-neutral [16, 75, 138], where domain expertise is often undervalued (e.g., in model building [110]). Furthermore, these automated tools are mostly evaluated with technical experts [4, 133, 134, 136, 143], with only a few recent empirical studies working with domain experts or identifying relevant values in specific cases [16] (e.g., values in computer vision datasets [113]). Therefore, we should find ways to engage those people who will use the technology (or be impacted by it) in values elicitation, and practice designing technology informed by those values [86].

Indeed, values play a central role in many of the professions with an increasing interest in automating data work, such as journalism newsrooms, legal research, and others [28, 48, 56, 69, 87, 90]. Similar to the expert data scientists mentioned above, domain experts can benefit from using these tools because they extensively rely on analyzing huge volumes of text documents [87] to improve production and attain high efficiency. At the same time, human values (e.g., transparency, justice [47, 109, 117]) are central to their day-to-day (text) data analysis practices. Therefore, it is important to understand not only their values expectations, i.e., the values

DIS '22, June 13–17, 2022, Virtual Event, Australia

ACM ISBN 978-1-4503-9358-4/22/06…$15.00
https://doi.org/10.1145/3532106.3533483

they perceive as relevant for automated text processing systems but also how those values should be encoded in system design [59, 79]. Elucidating such connections and tensions—between participants' values and NLP system design—constitutes an important area for research contributions. That said, the often tacit nature [cf. 55, 106] of these values can make simply eliciting them a non-trivial task [3, 139].

One method of eliciting values is through design fiction. Design fiction is well-known for engaging users in ethical and sensitive discussions about speculative technologies that do not exist yet [7, 89, 142]. Earlier work has applied this method for engaging a wide variety of potential users, such as AI engineers [72], students [139], and older adults [3]. The work presented in this paper was conducted as part of a larger project, where the research team is developing dedicated NLP tools for investigative data journalists and legal experts. Although prior work has described the general values to which certain professionals adhere [e.g., 42, 81], no existing work has empirically investigated domain experts' values expectations for automated NLP systems. The work presented here engaged directly with such domain experts, who have little to no formal training in computing, by using design fiction. Beyond pure novelty, this combination of methods and focus highlights the interplay between values and implementation in a way not seen in prior work.

The work presented in this paper mainly addresses the three following research questions (RQ's):
RQ1: *What values do domain experts such as journalists and legal experts care about from specific automated NLP futures designed for them?*
RQ2: *How do those values relate specifically with the technical implementation of automation involved in such systems?*
RQ3: *What are not-to-design implications, i.e., features, functionalities, or systems that should not be implemented?*

In addressing these research questions with our design fiction engagement, this paper makes three interrelated contributions. First, in response to ***RQ1***, we provide a detailed understanding of domain experts' values expectations and perspectives from automated language processing systems. Our findings indicate that journalists and legal experts deeply care about values of autonomy, freedom from biases, privacy, trust, and human welfare. Thus, automated language technologies (or any technology more broadly) designed for them, should adhere to these values in the design process. Second, our design fiction concepts, which were inspired and informed by initial exploratory interviews with journalists and legal professionals and relevant design futuring work [29, 65, 140, 142], provided a fictional world where participants could imagine that these automated NLP systems lived [40]. Doing so elicited reactions about the implementation of automation that, in response to ***RQ2*** were comparable to yet distinct from similar prior work in different (e.g., profession, user, and profit-oriented) contexts [16, 38, 49, 72, 134, 136, 142]. Third, in response to ***RQ3***, our findings further suggest design fiction has a unique capacity for enabling users to articulate not-to-design implications. Such implications emerged primarily around values mismatches between our participants and the design concepts. Although the study presented in this paper focused on automated NLP systems, this capacity of design fiction may apply more broadly. The paper also relates these results within prior work in language technology design, compares technical experts' and domain experts' perspectives on language automation, and provides a comparison of prior values work engaging users in other domains. Doing so illustrates how this work advances the conversation about methods for incorporating domain experts' values in language technology design.

2 RELATED WORK

This paper combined techniques from prior work in design futuring methods [82, 83] in participatory design workshop. We applied these techniques to engage domain experts from investigative journalism and legal analysis in conversation about their values expectations and perceptions of automated NLP futures.

2.1 User Groups and Domains Under Investigation

2.1.1 Investigative Data Journalism. Investigative journalists work with a wide range of datasets and data types. For instance, numerical tabular data such as Election Campaign data [69], text data from disparate text documents, such as thousands of pages of training reports, manual, court briefings, and record requests [42]. Our initial background interviews shows that, journalists apply various computational tools (e.g., Python, R, SQL) and algorithms to process numerical data [69, 108], and they process text documents almost manually. This can be a daunting and time consuming endeavor. Similarly, several research tried to address some these specific challenges with regards to text data processing, for example, Hassan et al. [71] studied presidential debates to uncover check worthy factual claims using ML algorithms. Jain et al. [73] applied a combination of ML and NLP techniques to support the investigative journalist by examining Connectas dataset [101]. There is also a growing trend toward building automated tools to support them [33]. While all these developments are already in progress, journalists values expectations from these automated tools are still an under-explored avenue [47, 91, 117]. This work aims to identify investigative journalists values expectations from several automated NLP design fiction concepts.

2.1.2 Legal Analysis. Legal experts also work with a wide variety of text documents consisting of case laws, legal articles, news, jury instructions from different jurisdiction among others [109]. Manual text analysis of these documents can be overwhelming for them, this was also confirmed from our initial interviews. Although there are tools out there to assist them (e.g., [107]), these tools are inefficient to support their specific needs. For instance, these tools may generate a large number of search results, sometimes they might display obsolete laws. In addition, interpreting the laws can be challenging because laws themselves are inherently ambiguous [109]. Recent research have addressed some of these specific problems, for example, Biagioli et al. [15] studied the provision classification using multi-class SVM model. Mok and Mok [95] classified sentences in breach of contract court decisions employing NLP techniques. Also, there is an emerging trend for building automated tools for the legal system [44].

Given the exponential growth in the development of automated language technologies both domain specific and domain agnostic [30, 35, 44, 135], we do not know how different domain experts will perceive them, most importantly – the "values" they care about or deem as important [59]. Therefore, as a case study, we wanted to bridge this gap, by engaging domain experts from the two above mentioned domains who extensively work with text data in automated NLP conversations. Doing so is essential because domain experts are often excluded from such conversations (e.g., [57, 110]), our research is an effort towards making that inclusion possible [39, 75]

2.2 Envisioning Automated NLP Futures Through Design Futuring Methods

Our design research exploration draws on various design futuring paradigms of critical design [8, 9], speculative design [53], and design fiction [20, 65, 66, 72, 139, 140, 142].

The purpose of critical design [8, 9, 54] is to challenge existing assumptions [53] by allowing the user to evaluate the designs to examine hidden dimensions such as inherent values, ethics, morals, underlying assumptions, believes, disbelieves, social values [139], cultural norms [9], and identify weakness. "A critical design should be demanding, challenging, and if it is going to raise awareness, do so for issues that are not already well known " [53, p.54]. The purpose of critical design is to engage users in the conversation or debate that is critical. It aims to suggest alternative design perspectives [9] by examining the here-and-now to imagine a future yet to exist. Critical design rejects the necessity of everything being nice, rather it welcomes dark designs, absurdity, humor, satire, able to invoke conflicting emotions in the viewer. One of the key feature of critical design is "provocativeness" [9], i.e., "A slight strangeness is the key – too weird and they are instantly dismissed, not strange enough and they're absorbed into everyday reality", [9, p. 294] [citing 54, p. 63].

Among other design futuring methods, Speculative design is about design imagination; this could be sociological, technical, dark and original imagination. Speculative design enables us to imagine future worlds shaped by our hopes, dreams, wishes and fears. Speculative design can "make a whole range of viable and not so viable possibilities tangible and available for consideration" [53, p.161]. In collaborative speculation or co-speculation participants are considered as a significant part of creative endeavor to generate ideas, question any assumptions, and imagine possible futures by engaging them in co-creation [43, 111]. Participants speculations are informed by their domain expertise, that we as researchers could not speculate without [132]. Notable work applying co-speculation includes Wakkary et al. [132] engaging philosophers, Desjardins et al. [43] engagement of users to speculate about IoT in home spaces.

Building off of these design approaches, we applied Design Fiction to build an automated NLP world for the journalists and legal experts, where our design concepts lived [40, 83]. "Design fiction is about creative provocation, raising questions, innovation, and exploration" [18]. Design fiction is useful during the ideation process and can be an essential tool for critical design [89]. While some design fiction is low-cost, creating quality design fiction can take significant amount of time [65] and effort [123]. It is an easy way to receive early critique of the designs by envisioning the future and it's impact [126, 139]. Several work has applied design fiction as the primary means of inquiry; for example, Diakopoulos and Johnson [49] applied fictional scenarios to examined the ethical implications caused by deepfakes by discussing possible harms and threats to voters, candidates, and campaigns; Houde et al. [72] also applied fictional scenarios to investigate the misuse of Generative AI by engaging AI engineers, Ballard et al. [7] and Wu et al. [142] engaged industry participants to raise their ethical awareness through games and design fiction respectively. As a speculative method, design fiction has been used as a value elicitation tool [72, 139] by foregrounding values [126].

3 METHODOLOGY

We conducted two participatory design fiction workshops at two non-profit organizations, one in the domain of investigative journalism and one in legal research. This section provides a brief background about each organization, our previous research engagements with these participants, the methods for creating the design fiction materials, and our data analysis methods.

3.1 Participant Background and Prior Research Engagement

This work was conducted as part of a larger project around designing text based tools to support domain experts' text data work practices. Our participants were professionals from two US-based non-profit organizations: People's Press (PP), an investigative journalism newsroom, and Justice Means Everyone (JME), a legal analysis group. [1].

3.1.1 Participant Recruitment and Demographics. We recruited participants via emailing participants from a previous interview study (described below). Four investigative journalists and four legal researchers agreed to participate in the workshop. Journalists' genders included three males and one female, and legal experts's genders included three females and one male. Investigative reporters' educational backgrounds were all BA or MA. Legal researchers' education ranged from Juris Doctor (JD) degree to BA or MA. Investigative reporter titles included Journalist Developer (P1), Algorithmic Reporter (P2), Editor (P3), and Data Journalist (P4). The legal professionals' titles included Head Attorney (J1), Senior Attorney (J2), and Attorney (J3, J4). Participants' technical experience varied widely. Most of the journalists had working experience (through self-taught) in programming tools such as Python/SQL [69], whereas the legal researchers had no programming experience, they relied on tools such as Google Docs [109] for text data analysis. All participants had no formal computational training. Participants were compensated with $30 gift card as a token of thanks. The study was approved by Lehigh University Institutional Review Board (IRB).

3.1.2 Prior Semi-Structured Interviews and Key Findings. In prior work, we conducted interviews with staff members at each of these non-profits. Those interviews involved nine participants in total, five investigative journalists and four legal researchers. Across the two sets of interviews, we asked similar questions and used a similar sketching exercise (via [125]) with slight modifications to fit the domain of interest. Sample questions can be found in Appendix

[1] All organizations, participant names, role titles are pseudonyms

A.1. All interviews were conducted between November 2019 and February 2020, and they generated 13 hours and 42 minutes of interview data in total. We analyzed the entire dataset using Grounded theory method (GTM) [34, 68, 128]. The full results are presented elsewhere [121] and summarized here.

The interviews showed comparable high-level data analysis practices across both domains. We also observed that both domains' data work practices required a certain level of manual human intervention. Regardless of the data types and computational tools applied across the domains, both groups of participants engaged in three high-level phases of data work practices: *Exploration*, *Verification*, and *Publication.* The *Exploration* phase involved collecting data, data transformation, and applying various tools and data analysis techniques to generate insights. The *Verification* phase involved fact-checking with internal/external sources to ensure correctness and accuracy. In the *Publication* phase, journalists publish news stories, and legal professionals publish legal reports. Across these phases, concerns and challenges included finding insights from data (e.g., manual process, time consuming), as well as reliance on multiple sources (e.g., subject matter experts) and technical processes (e.g., checking teammates work) to ensure accuracy. These data work practices and challenges inspired the automated text processing design fiction concepts described further below.

3.2 Participatory Design Fiction Workshop

Our primary engagement with participants combined the collaborative ideation and shared goals of participatory design workshops [98, 99] with the speculative nature of design fiction [18, 21, 53]. This combination of methods was used to explore participant's values expectations and perceptions about automated NLP systems. This subsection describes the processes by which the design fictions were created, the presentation format, and the manner in which they were used in our workshops. Throughout, we describe how our approach both draws on and extends prior work using speculative design methods, both generally [20, 29, 140] and specifically as a means of engaging with study participants [3, 22, 49, 139].

3.2.1 Construction of the Design Fiction Concepts and Visuals. One question we repeatedly asked during our design fiction construction process was: "how much of these design fictions are truly fiction?" Accordingly, the design fictions intentionally spanned from realistic and implementable to more speculative or fanciful. The overall sensibility combined a dash of modernism with a pinch of humanism, using vibrant bold colors [127], textures, and varied images. Similarly, typefaces were selected to balance familiarity and readability against speculative and unfamiliar futures. All depictions included details to help build a world in which these designs existed [40].For example, the DocSelfSorter (Figure 1, top left) was shown with a woman working in front of a computer and a "self-stirring" coffee mug on the table placed to make the concept feel simultaneously relatable yet foreign.

Inspired by Wong et al. [139], we initially depicted each design using several different approaches, such as Instagram/TikTok sponsored posts, FAQ pages, testimonial pages, landing pages, and Amazon product pages. We also presented our designs in varied contexts, such as on a desktop or a laptop, on a magazine page, and in the airport or highway billboards as adds. After careful evaluation and examination, we decided to present each design using two different formats (in two pages). The first page was either a landing page or a How It Works page from a product website (e.g., Figure 1, top middle). And the second was more varied, either a testimonial page, a FAQ page, an airport add, or an Amazon product page on a laptop, depending on the aspects of each design that we sought to foreground. We spent one and a half months in the summer of 2020 to create, curate, and refine our scrapbook of design concepts. We constructed these provocations using found images, collages, and overlays of images and text, inspired by both Blythe [19] and Gaver [65]. Although the design concepts varied in their degree of practicality, all of them were imbued with a touch of intentional ambiguity [66] to allow for multiple interpretations [116] and meanings [29]. This ambiguity was accomplished by leaving specific details about the systems' working procedures intentionally incomplete. On the other hand, "slight strangeness" [52] was implemented through raising conflicting emotions, that enable users to ask if the designs are real or not, serious or not, useful or harmful, and the designs placed participants in the position to judge for themselves [53].

These depictions were developed using a combination of PowerPoint, Balsamiq, and Picasa and other free online image editing tools. We intentionally avoided the application of powerful tools such as inVision/Sketch or Photoshop altogether, because we wanted our design fictions to be easy to make and sketch-like [31], while focusing on the quality of individual elements [65]. We took caution to ensure that our designs does not give the impression of a product specification or catalogue like [29]. Rather, we wanted to create evocative images that gestured not only at the design itself and its technical functioning but also at the broader future world in which they exists [40]. Similar to diegetic prototypes, "these technologies only exist in the fictional world — what film scholars call the diegesis — but they exist as fully functioning objects in that world" [77, p. 41].

Informed by our initial interviews (described in 3.1.2), we designed and organized our design fiction concepts around the three phases: *Exploration, Verification*, and *Publication* [37, 121]. Due to high-level similarities across both participant domains' data work practices, primary necessities, and challenges, similar designs were used across both groups of participants. The designs were intentionally crafted to provoke critical discussion about issues of values, ethics, ideology, and design assumptions [8, 9, 115]. At the same time, the designs also draw inspiration from current NLP research [6, 11, 46, 129, 137]. Unlike other research using design fiction's [3, 22, 140], this grounding in the related technical literature provides both a sense of feasibility and gives our participants additional details to which to attend. We were also sensitive to our participants' wants and needs, while making sure to limit researcher's influence in the designs when framing and shaping ideas [3, 19]. During this process, many of our initial design concepts, such as automated writing tools or conversational agents, were abandoned, either because something quite similar already exists, because the concept was too far removed from current NLP technologies [14], because the concept was too mundane, or because they were not in alignment with findings highlighted in our initial exploratory interviews.

3.2.2 Description of Design Fiction Materials: A Computational Scrapbook From the Future. In addition to being aware of participants' wants and needs posed during initial interviews, we also brainstormed by reviewing related literature about the latest tools in investigative journalism [48] and legal research [44], as well as literature about current NLP capabilities for ideas and inspiration [30, 46, 88]. This step resulted in many initial and sometimes irrelevant design concepts (e.g., automated elevator showing news). To further narrow down, and specifically focus on participants data work practices' and challenges, we organized these design concepts around three phase of *Exploration*, *Verification*, and *Publication* described in section 3.1.2. This step led us to the following design concepts across each phase: 1) *Exploration*: DocSelfSorter, DataWatch; 2) *Verification*: VeriFact, and TransparencyMonitor/CompassionMonitor; 3) *Publication*: VRLegalPal/VRNewsPal. Interestingly, all these designs had one thing in common, NLP automation [30, 35, 46]. NLP automation heavy futures were created by the challenges and the primary necessities described in the initial exploratory interviews. We were not interested in engaging participants in the detailed underlying workings of NLP, rather, we were interested to provoke a reaction by engaging participants in a discussion and conversation about NLP automated tools to learn about their values expectations and perceptions [136].

Our final scrapbook had five design concepts, and was presented in the following sequence: DocSelfSorter, DataWatch, VRLegalPal/VRNewsPal, VeriFact, and CompassionMonitor/TransparencyMonitor. Two designs had different titles to reflect professional domain and organizational values, they are VRNewsPal/VRLegalPal and CompassionMonitor/ TransparencyMonitor, respectively. We decided to show the designs in this sequence based on participants' original data work flow, as well as moving from more practical to more speculative and provocative [8]. A brief description for each of these design fiction concept is provided below along with detailed data work challenges and needs.

Design Fiction Concepts in the Exploration Phase: The DocSelfSorter shown in Figure 2 was based on the needs and challenges participants described for analysis of text documents. Such analysis include automatically sorting similar articles, discovering important topics, text classification, framing in the input documents, etc. [6, 11, 94, 129]. This concept was presented as How It Works page, and a testimonial page from the perspective of a student and a professional journalists/legal personnel. For this design concept, we added the automation aspect by stating "No manual human intervention required" in the How It Works page. Whether the concept represented a software/hardware and its algorithmic details was left open to participant interpretation [64]. This concept had little to no variation across both domains.

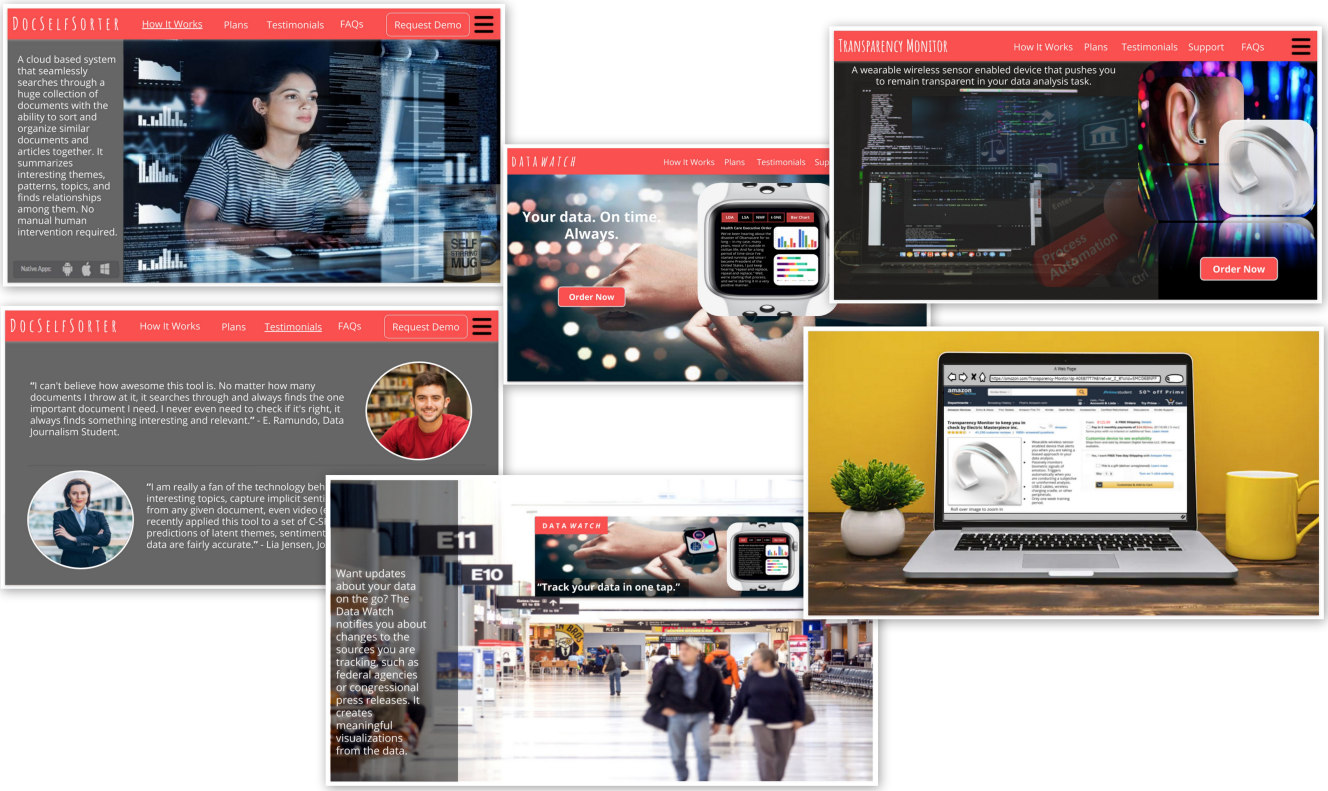
The next design concept in the scrapbook was the DataWatch (in Figure 4) inspired by both organizations need to rely on tracking various sites such as Social Networking Sites (SNS) sites such as politicians' twitter/Instagram, and legal sources (e.g., case laws, federal law updates). We also added the capability to perform simple data analysis in the watch platform. Data source tracking and simple visualizations capabilities were embedded in a smart watch interface. The concept was presented as a landing page and an advertisement at the airport billboard [139, 140]; we portrayed (the airport advertisement) the watch to analyze both numbers and text data to bring small "ambiguity of information" described in Gaver et al. [66] study. Even though the concept (landing page) showed text analysis using various Topic Modeling algorithms [129], however, technical details was not included to allow open interpretation [64]

Design Fiction Concepts in the Verification Phase: The third concept, VeriFact, was based on both organizations need for validating/verifying the veracity of the data (e.g., text, numbers) analysis (see Figure 6). Fact checking is a common practice in both organizations but the meaning and how it is done varies widely. For instance, data journalists check facts/data in various ways, for example, comparing teammates data analysis output with their own [108] using tools like excel/python, whereas legal researchers check each others text (e.g., statue, case law) classification task manually; both practices involves communication and collaboration, and manual human-intervention [108, 109]. The automated fact checking aspect was embedded in the landing page tagline "Taking guesswork out of fact checking". This concept was presented as FAQ page including questions, as a way to invoke provocativeness [8]. The FAQ page emphasized algorithms, common errors, privacy aspects of the design. An example question was "How does VeriFact ensure the anonymity of my sources?" These FAQ questions were modified as per domain requirement.

Design Fiction Concepts in the Publication Phase: The VRLegalPal/VRNewsPal was one of the most futuristic speculative design concepts, inspired from the latest development and demand in the virtual reality [108] immersive technology (see Figure 8). The VR system was able to reconstruct VR scene/stories from courtroom trials or news. This design was presented as a landing page followed by user testimonials (e.g., journalism student, attorney). The landing page included the tagline "Walk in their shoes, See through their eyes ..." to allow the users to evaluate the design from different different users perspective, to induce emotional response among the participants [32]. The testimonial's conveyed the idea of how this design have helped different people to understand difficult terms/concept, also to establish empathy towards different subjects (see Figure 10).

The two design concepts specifically designed using organizations core values were CompassionMonitor and TransparencyMonitor, i.e., compassion for the legal researchers, and transparency for the journalists. These concepts were presented as wearable tech jewelry (e.g., bracelet for hands or for ear jewelry) [140]. They utilize sensing technologies [139, 140]. These systems contain biometric sensors that enables the user to remain compassionate (e.g., transparency for People's Press) in their data analysis approaches. The design was presented as an Amazon product page [139] on a laptop, the product details indicated high level technical details such as training period, USB cables.

These five designs were wrapped inside of front/back cover of a scrapbook (e.g., "memory scrapbook" [86]). The front cover of the scrapbook contained the heading "Curated Collection of Speculative Computational Objects", with a sub-heading "... a future scrapbook... ", to clearly demonstrate the speculative aspect of the designs [53] Each design was followed by a prompt asking "what do you think?", then "In my opinion, the design is ..." on a blank page. The scrapbook



A cloud based system that seamlessly searches through a huge collection of documents with the ability to sort and organize similar documents and articles together. It summarizes interesting themes, patterns, topics, and finds relationships among them. No manual human intervention required.

“I can't believe how awesome this tool is. No matter how many documents I throw at it, it searches through and always finds the one important document I need. I never even need to check if it's right, it always finds something interesting and relevant.” - E. Ramundo, Data Journalism Student.

“I am really a fan of the technology beh
interesting topics, capture implicit senti
from any given document, even video (e
recently applied this tool to a set of C-SI
predictions of latent themes, sentiment
data are fairly accurate.” - Lia Jensen, Jo

Your data. On time. Always.

A wearable wireless sensor enabled device that pushes you to remain transparent in your data analysis task.

“Track your data in one tap.”

Want updates about your data on the go? The Data Watch notifies you about changes to the sources you are tracking, such as federal agencies or congressional press releases. It creates meaningful visualizations from the data.

Figure 1: Three speculative design fiction provocations represented as cards from the future; from left DocSelfSorter (top-left as a How it Works page, left-middle a testimonial page), DataWatch (middle as a lading page and on a billboard add on the airport), and Transparency Monitor (top-right as a landing page right-middle as an amazon add on a laptop). For a more closer look see Appendix B.

also had an ending prompt of “What does it feel like to live in a world with these technologies?” Like the front cover, the back cover also used headings “ Legal research technology from the future”, subheading “A Computational World Imagined ...”, to illustrate that these designs are conceptual and for speculation purpose only. The purpose of these prompts was to provide the users the freedom to think, instead of guiding them to think in a predefined manner.

3.2.3 Participatory Workshop. The first workshop was conducted with legal research professionals in July 2020, and the second workshop was conducted with expert investigative journalists in August 2020. Both authors actively engaged participants during the workshop.

Before The Workshop: Once our workshop session was scheduled, we emailed them the design concepts in a single PDF file one week prior to the workshop. In the email, we explicitly mentioned about the speculative nature of the design concepts. Although we asked our participants to answer (through PDF annotation) the prompts in the provocations, doing so was not mandatory. The online nature of our engagement (described more in the next paragraph) was more conductive to presenting our scrapbook as a PDF, which is also portable and allow easy annotation.

Workshop Session: The workshops were conducted online via Zoom, since the COVID-19 pandemic precluded the possibility of conducting an in-person workshop. Each workshop was audio recorded, generating 3 hours and 8 minutes of data in total. The workshops began with a brief description of the goals of the session, as well as a brief description of the relationships among four different types of futures: possible, plausible, probably, and preferable [53, 131]. The researchers then presented the provocations in the form of a presentation one after another, followed each with questions and discussion. A sample set of questions can be found in the Appendix A.2. We followed a semi-structured protocol, asking a similar set of questions for each of the provocations, as well as asking follow-up questions on the fly. As noted above, we conducted the first workshop with four legal researcher’s from Justice Means Everyone, and we conducted the second workshop with four investigative journalists from People’s Press.

3.3 Workshop Data Analysis and Coding

The workshop session was audio recorded, anonymized, and transcribed (using [45]) for data analysis. We applied Thematic Analysis

(TA) [5, 25, 26] to qualitatively analyze audio transcript data. Our research question(s), data collection informed TA as a suitable method for data analysis. The codes, themes, and categories emerged inductively from the data. In that way, the emergent themes speaks and remain close to the data. The process was also deductive, when we analyzed data from Value Sensitive Design (VSD) framework analytical lens [59] to uncover participants values (e.g., inductive and deductive [16]). We analyzed entire collection of data. The first author performed thematic coding, and discussed the emergent codes and themes with the second author, whenever there was a conflict or a new theme emerged, the theme was reported only after reaching congruence. We iteratively analyzed the data beginning with the first design concept (DocSelfSorter) across both workshops, then the second concept (DataWatch) across both workshops, and so on until the fifth design concept. We also revisited data (multiple times) for clarity, further insights and understanding whenever required. The themes obtained from the data were compared and validated across both groups [25, 26]; we were also sensitive towards surprising themes and findings

4 RESULTS: VALUES EXPECTATIONS, AUTOMATION, AND WHAT NOT TO DESIGN

Participants' responses to the design concepts were similar and comparable across both organizations. These responses are organized here in terms of three high-level, non-exclusive themes. First, the design fiction concepts prompted participant's to express their *Values Expectations*, i.e., the values that participants perceived as most important or salient, es well as how those values were implemented (or not) in the design concepts. We analyze these values expectations from the perspective of Value Sensitive Design (VSD) [58–62]. Second, participants' responses often mentioned *Tensions with Automation*. Such tensions emerged between their values expectations and the kinds of automation suggested in the speculative designs. Third, such tensions highlighted *Not-To-Design Implications*, i.e., the qualities that automated NLP technologies should *not* have. Although described separately here in the interest of analytic clarity, discussions of these various themes and concerns were highly interwoven during the workshops.

4.1 Domain Experts' Values Expectations

As much as our participants were excited to talk about their hopes and wishes about the design fictions potential to support their data-driven work practices, they were also very much anxious about the power that each of those designs carries that could inadvertently be misused and cause harm to the humans. We categorized participants' responses into five major themes, each associated with a value identified in the previous work [58–62]: *Autonomy and Subjectivity*, *Biases*, *Invasion of Privacy*, *Trust*, and *Human Welfare*.

4.1.1 Autonomy and Subjectivity. Participants frequently discussed the design fiction concepts in terms of human-machine configurations [cf. 124]. That is, discussions focused less on the abilities of a given technology *per se* and more on configuring the relationships between these technologies and humans. For instance, the DocSelfSorter often elicited comments about participants' desire to have some kind of human control in the system, similar to the concept of an user autonomy [58, 62]. Autonomy is described as *"individuals who are self-determining: who are able to decide, plan, and act in ways that they believe will help them to achieve their goals and promote their values"* [62, p.2]. A legal researcher, P04, described the complexity associated with manual human text classification (e.g., legal terms in statutes across various US states) to illustrate this.

> *"I think I'm just like always a little bit skeptical of technology. [...] But I do think that there will be like some human element of state-by-state interpretation, no matter how advanced the system is"* - J4, Attorney.

In their work, J4 and her team had to look beyond the legal statutes to understand the proper meaning and interpretation of each term in the context of different states; the entire process was time consuming and laborious. Put differently, even if a system could perform text classification with ease, the system may not be able to navigate the subtle differences between the how the laws are written and real-world legal practice. In such tasks, human interpretation and understanding of laws are always necessary.

Furthermore, in addition to their years of experience in the legal system, they sought out external legal experts [109] to confirm the validity of their text classification. J1 expressed her desire to have such capabilities implemented in the DocSelfSorter design itself, saying, *"I would want to have the ability to confirm with the human expert [...] because then we could easily go to our people to say, this is why we think X in the state. Is there anything that gives you pause about this conclusion, then I could confirm"*. Again, emphasizing human interpretation and human assessment for this task [136].

Human autonomy was also important from the perspective of participants' professional practice. For instance, legal expert P03 described that the lawyers make subjective assessments, including using their judgment, combining evidence and intuition, and learning from trial and error. Since the DocSelfSorter was an automated system by design, it was unable to replicate these human abilities, making it less efficient and somewhat impractical for real-world application. Reacting to the design, J3 felt

> *"In one sense comfortable because, we don't have to worry about a human screwing it up. But on the other hand, that's always the risk, but that human judgment, always sort of [...] something looks like a big risk [...] some level of human intuition play and the decision on the talents. And yet that doesn't always work, but it's better than the alternatives"* - J3, Attorney.

In this quote, we see J3 was wrestling with the fact that human subjective judgment is fallible. At the same time, though, J3 embraces the importance of human intuition. That is, this quote exemplifies participants' perception of tradeoffs inherent in the tension between subjectivity and automation.

Journalists also expressed a similar desire to have some human autonomy in the system as the way of human-enhancing systems. For instance, journalist P4 insisted that the DocSelfSorter should allow the humans to do things they are currently unable to do by providing some form of human control [58, 62] in the system. Participants relate the DocSelfSorter to support them in generating story ideas [108], and for that they imagined whether the system could provide them the ability to ask complex questions or allow them to add additional datasets to generate story ideas. P4 further emphasized that without such autonomy, a system can practically become not so useful. He provided the following example,

> *"[...] like something that will allow a human to reasonably go through a task in the amount of time needed to produce a story [...] Like, if you just throw documents into a system and ask it to classify, you may find that like, a good example of something going awry, like X tried to like do a thing to forecast, Bill's passing in Y. And it turned out that their model was just had taken all the documents from every bill and every step in the process. [...] It was completely useless in real time. [...] we probably need to think about like* ***systems that are not going to operate independently of humans****, but which are going to allow humans to do things that they can't do"* (emphasis added) - P4, Data Journalist.

These excerpts illustrate that these systems are not expected to work without human. Rather, they should work with the human to support them and enhance their current abilities, by providing the right amount of control at the right time.

Although the above focuses primarily on DocSelfSorter, not allowing for users' autonomy and the exercise of participants' subjective judgement proved important across reactions to multiple design fiction concepts, essentially, for achieving goals and promoting user values in their professional practice [58]. Both groups of participants exemplified how their autonomy was violated, while also suggesting how their autonomy could have been protected by incorporating certain necessary capabilities into the designs. As noted above by P3, increasing support for autonomy and subjectivity in a system may simultaneously also increase the possibility of introducing bias and other potential misuse [59, 61].

4.1.2 Freedom from Biases. Bias as defined for computer systems by Friedman and Nissenbaum [61] means *"systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" [61, p.3]*. Both groups of participants described the importance of avoiding different biases in their current professional practices. For example, journalists expressed how the current technologies allow them to navigate the issues of biases [59, 61], and how they are able to consider specific context associated with a situation and practice awareness. However, automated tools like DocSelfSorter were perceived as prohibiting them from such practices. According to Developer Journalist, P1,

> *"There's this balance that exists now, between sort of like, human judgment and context and biases, and sort of habit [...] versus like the, the capabilities, the technical capabilities of computational systems. [...] But I do think that like an ability to sort of guide and manage, like building that into systems into these kind of technologies, I think like would make it at least something that like, [...] I'd be able to sort of come to grips with, or at least make a rational decision about rather than an irrational sort of like, no way, forget it kind of scenario. [...]"* - P1, Developer Journalist.

Similarly, when the system design constraints the users to use a system by not providing the ability of practicing human judgement, interpretation, decision making support will introduce technical biases [61]. It will also hinder human autonomy [62], illustrating how these values are in practice interwoven.

As an alternative strategy, the legal professionals work around issues of bias by remaining neutral about revealing demographics information. This strategy, spanning both arguments made in the court and their published written reports, aims to avoid pre-existing biases associated with demographics (e.g., race, gender) [58]. Drawing on this, J1 pointed out an important design flaw in the VRLegalPal, noting that *"when you're having virtual reality, if there are then people in there, we just want to make sure we're not reinforcing anything"* in terms of those biases. J1 further emphasized researching relevant sources (e.g., social science, medicine) to consider issues of implicit and unconscious biases.

Participants also envisioned ways that issues of biases could creep into the designs when they described using the designs from various stakeholder (direct/indirect/excluded user) perspectives. For example, legal expert, J2 described the potential for mishandling if VeriFact were used by a judge.

> *"Like a judge saying, [...] I want to bypass what the attorney's had to say, and I just want the output from the system. I just want to see what the system says. I don't want to hear your arguments because this system can produce the argument, the counter argument. I can read it for myself. That saves me more time. So, maybe it becoming a tool with the unintended consequence of bypassing the justice component of the system"* - J2, Senior Attorney.

Indeed, this account reflects the ways that risk assessment systems have been adopted in some courtrooms [27, 38, 74]. As J2 highlights, a mismatch between who the system is designed for, and who might use it, can introduce biases (e.g., emergent bias) [13].

These examples illustrate how participants linked aspects of technical specifications – e.g., degrees or types of automation, presentation of demographics in virtual reality – with the value of avoiding bias [61]. The design fiction concepts also enabled participants in envisioning the system being used by a different group of user than for whom it was designed, an envisioning which highlighted yet another potential for bias. Such shifts of context – considering a design from a different perspective or being employed by different users – also highlighted concerns about privacy.

4.1.3 Protection of Privacy. Privacy as a value *"refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others" [59, p.22]*. The design fiction provocations raised privacy concerns [cf. 59, 139] most often when participants placed the designs in different contexts or considered them using from different stakeholder perspectives. These perspectives included not only participants themselves but also other kinds of relationships with the fictional systems [10]. For example, J1 talked about how the VRLegalPal might expose relevant information, because it reconstructs a VR scene from transcripts, which might not be redacted. She said,

> *"I was at first thinking on this, like privacy [...] without fully thinking through the ramifications where now defendants have access to the court rooms, [...] privacy, even if it's in a survivor's home, unless they put up a protection order. So there was a piece of it that was like, this could maybe help preserve against that. But then I was like, wow, this could really be an invasion of privacy now"* - J1, Attorney Head.

J1 went on to describe how such a system might violate the privacy not only of the victim/survivor, but also of the defendant.

In contrast, the journalists' expression of privacy concern was more toward themselves. For example, journalists P1, P4 collaboratively discussed privacy issues with TransparencyMonitor, he provided the context of its use in their job, he said,

> *"this should not be able to connect to the internet in any way [...] I think like, essentially it almost becomes a question of like, who has access to the output of it, the product governance, you know what it's saying? Like, as an employee, does that factor into an evaluation that I got, [...] I mean, maybe should, but essentially like what portions of that, like if I took corrective action because of it, like, do you use that, does that factor in [...]"* - P1, P4, Journalist team.

These excerpts illustrate how the design fiction provocations encouraged participants to consider multiple perspectives, not only that of direct users [10], and how those multiple perspectives surfaced privacy concerns. Furthermore, these specific concerns about privacy violations also introduced a more general lack of trust in the systems [59, 60].

4.1.4 Trust. Participants repeatedly raised questions about not being able to trust these automated systems [59]. According to Friedman et al. [60], *"we trust when we are vulnerable to harm from others yet believe these others would not harm us even though they could"* [60, p.2]. How could they trust the CompassionMonitor or TransparencyMonitor that would actually monitor the attitudes they claimed? How could they be certain that DocSelfSorter would find the patterns or themes they wanted it to? These concerns were particularly salient with VeriFact's automated verification of facts in written reports. Both group of participants already engage in rigorous verification processes in their respective professional domains [81, 109], and they reacted to the design through the lens of these existing practices. For example, news Editor P3 reacted to VeriFact instantly by saying she would not trust the system without knowing how it works. Although she might use the system as a back up, so long as the system is easy to run, P3 still emphasized the importance of her own judgement over an automated system. If VeriFact rejected a claim she was claiming to be true, she said

> *"I wouldn't diagnose the system. I'd just be like, okay, let me go back to where I got this originally. [...] unless I am just misinterpreting something completely wrong, I'm going to trust myself and my judgment over kind of a black box telling me that something's incorrect. [...] I will fact check or check things compulsively. [...] if I feel satisfied by the facts that I've found, and I think that I am right, I would override, a program telling me that I'm incorrect"* - P3, Editor.

This example illustrates the interplay between the value of trust and the value of autonomy, as described above. As another journalist put it, *"I think you're gonna ask a bunch of journalists, like whether essentially, [...] how far we can trust automation. And I think most of us are going to be like, well, without me being there, not that much"*. This reaction is in alignment with expert data scientists response of trusting manually-crafted low-performance models over automated data science tools generated high-performance models [133]. Verification of facts is an essential journalistic practice, in part because it relates closely with the core value of conveying truth to the audience [81]. Trusting an automated tool for such an important task was thus perceived as abdicating professional responsibility.

In comparison, legal experts raised more concerns about potential misuse of VeriFact. For example, legal expert P03 imagined the design being used in a courtroom. He feared that the presence of VeriFact might hinder legal professional practice by limiting jurors' viewpoints and by not making *"subjective credibility assessments."* In J3's own words *"I think this would be an incredibly dangerous tool"*.

These excerpts illustrate how participants' trust in a system was linked with the degree to which the design preserved their ability to make subjective judgments. VeriFact directly conflicted this value of trust by simply telling users what is correct/incorrect. On one hand, providing more information about analytic details might increase participants' willingness to trust such a system. On the other, adding such a feature would fail to address the fundamnetal values misalignment between using an automated system for fact-checking and trustworthy performance of one's professional duties. Going further, the above comment about VeriFact being *"an incredibly dangerous tool"* goes beyond not trusting the system entirely to highlighting fears of possible misuse and future harm, described next.

4.1.5 Human Welfare. Across multiple design fiction concepts, both groups of participants also discussed the importance of establishing truth, as a way to establish human welfare which *"refers to people's physical, material, and psychological well-being" [59, p.22]*. For example, in the DocSelfSorter design, journalist P2 talked about his fear that people may not consider other analytic perspectives when they have specialized automated tools like the DocSelfSorter. People might take this system output as standard without being aware of the fact that a single model is unable to capture all meta-data (e.g., demographics, background or past records) information in the analysis. P2 went on saying,

> *"The thing I'm more concerned about is, [...] you use certain methods to analyze and it's sort of like* ***becomes the truth*** *about it, right. And even though, you know, there's, there's multiple tools and angles that have been used to look at it. [...] whatever model you choose to use, [...] like there's a danger of it"* (emphasis added) - P2, Algorithmic Journalist.

Likewise, legal professionals expressed a similar concern. For example, J1 feared that if a judge takes the automated system output as "truth," and what if they ignore other opinion or perspective. In worst case, they might assume that *"this is based on a perfect algorithm"*. J3 further expanded on this saying,

> *"My worry for this would be potentially is sort of the results drawn from the [DocSelfSorter] blows people's minds off to other authorities [...] other arguments, and be a weapon [...]"* - J3, Attorney.

To summarize, a system that automatically generates output without providing much details about how those results were generated and does not allow the user to have any say in the final outcome is an absolutely undesirable system, since they violate several values integral to both journalism and legal professional domain, even though these domain users may not necessarily be formally trained experts in computational systems. These violations naturally led us to users perceptions about automated NLP futures.

4.2 Tensions with Automation

Even though our design concepts were fictional and in some cases highly speculative, our participants easily envisioned themselves using and living with the design concepts. In doing so, a recurrent concern was automation, both how much and what kinds of automation were desired. Although closely related with the value of autonomy (Section 4.1.1), automation often raised cross-cutting concerns that went beyond any single value.

As noted above, complete or full technology automation was highly undesirable, even to participants lacking formal training in algorithmic tools. Journalists and legal professionals frequently make subjective judgments and contextual inferences based on available evidence [2, 51, 81, 105, 109]. For example, as we saw earlier, the journalists make subjective assessments to make informed decisions around issues of biases [58, 61] (among other values), since *"journalism's first obligation is to the truth"* [81, p.36]. Similarly, for the legal practitioners, *"justice sometimes requires an ability to bend the rules"* [56, p.51], as demonstrated in J3's sentiment about how his compassion is victim focused. Full technology automation, it was noted, would violates that autonomy [58] because it gives users fewer options to achieve their goals and promote their values in their professional work settings.

This attitude can be seen, for instance, in responses to the DocSelfSorter. This tool was not a viable option for the journalists and legal experts as it would get in the way of how they currently approach their text data work, which is thorough, rigorous, and involves multi-step decision-making processes. For example, legal professional J4 described that a large part of her document related work involves *sorting* and *analysis.* For her, sorting involves organizing similar documents together, while analysis involves classifying laws. J4 described how she and her team completed the painstaking task of reading statutes on a state-by-state basis to classify which statutes pertained to certain terms (e.g., force, consent). At the same time, this classification process involved significant analysis and interpretation. Some of these legal terms are inherently ambiguous [109]. Thus, J4 and her team had to consider additional sources (e.g., case laws) to avoid misinterpretation and fully understand the meaning of those terms in practice. In J4 own words,

> *" [...] what a statute may have said on its face was not always, what the law ended up being, because you have to really like analyze the statute in relation to case law and jury instructions and kind of put everything together. And it was just like this **multi-step analysis**"* (emphasis added) - J4, Attorney.

Thus, not only would fully automated systems be difficult to fit into their current text data work practices, such automation is also not what they even desire [134].

At the same time, neither journalist or legal experts actually desired absolutely no automation. Participants voiced desires that NLP systems be fast, accurate, and easy to use, as long as such systems could explain how they work. For example, for fully automated tools like VeriFact, news editor P3 said,

> *"I could see using a system like that as a backup, but I'd still do everything I do by hand anyway. [...] **I just wouldn't like trust, something that important to something that I can't see how it works or, I would just end up doing everything by myself**. [...] like if it's easy to run and maybe it would catch something that I missed anyway. [...] I wouldn't rely on something as important as fact checking or data checking as my sole thing"* (emphasis added) - P3, Editor.

This excerpt illustrates that, beyond just different levels of automation [cf. 134] each data analysis phase (exploration, verification, publication) might require a different kind of automation support, depending how important that step is. A similar need to describe how and why something works was expressed by journalist P4, not only for themselves, but also for the audience:

> *"[...] about what it's trained on [...]. I mean, you get really different results when you train a thing on with different questions in mind. I think they [the audience] would like both what it's trained on and then confidence intervals [...] I think like showing that to the public is helpful"* - P4, Data Journalist.

These excerpts illustrate the importance of data and algorithmic transparency [47, 50] in journalistic practices for automated NLP systems [133].

Thus, neither full automation nor no automation were seen as viable options. Instead, both groups of participants recounted their desire to have a "Human-in-the-loop" system. Similarly, in the above values analysis, many values violations were related to designs that emphasized traditional ML values—automation, efficiency, performance, etc. [16, 133]—over human values. For example, in case of DocSelfSorter, a legal expert J4 shared her perspective that any advanced technology will still require human interpretation and understanding. J1 further corroborated that point by asking for the ability to reject/confirm the system's prediction based on their own experience and subject matter expert (SME) (e.g., researcher, academics) suggestions. Similarly, P4 wanted human enhancing systems, rather than human replacing systems. While discussing the process of finding interesting story ideas, he said,

> *"I think like **human in the loop in some ways it's like is the value system in the loop**. I mean, it's hard to get away. All of these questions seem to touch ultimately on like a set of values goes into the picture somewhere"* (emphasis added) - P4, Data Journalist.

This approach our participants advocated could thus be described as "System-suggested Human-directed" automation, similar to expert data scientists' stance toward AutoAI/AutoML tools [134, 136]. Doing so was seen as a means to balance between the undesirable extremes of full automation and no automation.

4.3 Not-To-Design Implications

The above subsections discuss a variety of ways in which automated NLP technologies presented in our design fiction concepts were seen as unable to protect, or even at times violated, the values important to journalists and legal experts. Rather than envisioning entirely dystopian or post-apocalyptic scenarios [22, 72, 126], our participants instead focused on scenarios that were more quotidian but still just as undesirable or, as participants often put it, simply "inappropriate." This subsection organizes and analyzes these imagined scenarios, borrowing the conceptual framework from Baumer and Silberman [12] to offer implications about NLP technologies not to design.

4.3.1 Low-tech, No-tech, or Something Else? As described above, our participants' current text processing activities are complex, made up of both technical and manual human interventions. However, not only does this messy text data processing practice work for them, but a non-technological element is in fact necessary to these practices. For instance, for the legal experts, the terms in the documents themselves are often ambiguous [109] and require human interpretation. Furthermore, the correct classification of a given document depends not only on the content of that document but also on how other documents are interpreted and applied to the classification task. J1 provided an example, indicating the level of sophistication the automated tool might need to have in order to be able to truly support their work,

> *"[...] the ability to push back on what is known based on other things that are known that can push us towards the right answer"* - J1, Attorney Head.

Thus, these legal analysts consider not only what arguments *have* been made but also potential arguments that *could* be made. Their current, mostly manual text classification processes, although tedious and time consuming, allows them to navigate such situations in ways they suggested as difficult, perhaps impossible, to automate. Rather, the design challenge becomes how to provide the kinds of technological interventions that participants desire while simultaneously retaining the valuable aspects of their current *low-tech* approaches.

On the other hand, the journalists current data work in news storytelling involves a combination of computational tools (e.g., sql, python), human assessment, judgement and ethical decision making [47, 117]. Sometimes, they also engage the audience for data interpretation, as illustrated by journalist P1,

> *"When we got like [anonymized source's] emails and like what we did there was we literally just published them all and then invited people to tell us what was interesting, which is, [...] sometimes useful, but also weird and slow and not particularly accurate"* - P1, Developer Journalist.

This except illustrates how journalists combine computational tools and human judgement, potentially including the audience's. These capacities would be difficult to replaced with automated tools. Even if they are sometimes *"weird and slow and not particularly accurate,"* they are also *"sometimes useful"*. Put differently, even with a suite of automated tools available, journalists would likely still employ low-tech methods such as crowdsourcing.

4.3.2 Harms Outweigh Benefits? Participants mentioned a variety of ways that probable harms from design fiction concepts likely outweighed the potential benefits [59].

Some examples involved unintended use by those who are not perceived as the primary user of these automated tools. For instance, with automated language technologies, it is critical to provide sufficient explanation about training data: annotator/collector demographics, procedures and instructors for annotators, etc. [cf. 67]. Even if it improves efficiency or provides novel insights [16], a system that is trained with one demographic and is tested on a different demographic can perpetuate, reproduce, or even enhance biased outcomes [13, 14, 120].

Another important consideration about unintended harm is determined by who else have access to these technologies, in particular, if the designs specifically crafted for journalist and legal experts are used by other unintended stakeholders. For example, legal experts shared their concern if the VeriFact design was used by judge or a defendant (e.g., similar to bad actors taking advantage of large language models [14]). More precisely, VeriFact can be misused by a judge, if they start blindly believing the system outcome as truth, and by completely ignoring attorneys perspective, what if they "became slave to the task system" [56, p.41]. Another example of harm can occur in case of VRLegalPal, where the defendants will have access to information such as victim's home. These examples illustrate, how depending on who have access to these automated systems can have a huge impact how automated tools can become really dangerous.

Participants also mentioned fears about their job security due to these specialized automated tools. For instance, in J1's own words, *"job security would be number one, [if these tools can] do it all, why are we here?"*. The journalists were also vocal about it, however, their stance went from logic to philosophical musings, they further generalized their lack of trust in these automated systems. Tools like TransparencyMonitor made them really uncomfortable, almost pushed them to the edge, where they collectively said that, *"this should not be able to connect to the internet in any way [...] Like as an employee, like, does that factor into an evaluation"*. In addition to privacy concerns [58], these sentiments also echo several real life consequences for human welfare [59], such as public service workers who have lost their jobs [56, 103, 122], or users who went through inhuman experiences (e.g., complaints went to an automated chatbot, medicaid/welfare application denied unethically) due to automated tools.

Most importantly, the inability to *"guide and manage, like building that into systems into these kind of technologies"*, made the journalists not trusting those systems. Rather, they became frustrated, and went on saying,

> *"[..] I think it would be like deeply frightening if tomorrow [I] woke up and these sorts of things exist. [...] I'm*

> *just going to unplug all the cables in the house and, you know, move to the woods"* - P1, Developer Journalist.

Such statements illustrate the degree of participants' concerns for human autonomy [58] and the reactions when this value was threatened.

Collectively, such examples reinforce that not-to-design implications arise in large part due to values tensions. That is, such instances often occur when the values embedded in a design – e.g., data capture, accuracy, efficiency, or automation [16] – differed from the professional or personal values of participants – e.g., privacy [58], freedom from bias [61], human welfare [59], or human autonomy [16, 58, 86]. Thus, despite potential benefits, participants were generally reluctant to embrace fully automated systems.

4.4 Findings Summary

Overall, our design fiction workshop brought difficult but important technology conversations [53]. The above findings illustrates that our design fictions engaged participants in values elicitation with respect to automation and the various ways in which speculative automated NLP designs were inappropriate. These findings were expressed when both user groups perceived, evaluated, and reflected upon situating the design fiction concepts in a wide variety of contexts involving both direct users (e.g., journalist, legal experts) and other indirect or even excluded users (e.g., defendant, employer) perspectives, and whenever participants values of user autonomy, trust, invasion of privacy, freedom from biases, human welfare [58–61] were in conflict with the values encoded in the automated designs (e.g., efficiency, automation, performance [16, 142]).

Our participants did not appreciate fully or even mostly automated language systems. This reaction occurred regardless of their formal technical training/background, and despite the fact that the designs were dedicated to reduce some of the heavy burden of their human-labor intensive, time consuming data work practices. Instead of *full/complete* automation or *no* automation, participants desired human-in-the-loop (human guided) types of automation [134, 136]. Several examples revealed that journalist and legal professionals often, perhaps always, take a subjective stance based on available evidence and context [2, 51, 105]. Such subjectivity is where humans excel but the machine cannot. These practices of subjective assessment, intuition [102], judgment, empathy help participants navigate issues around biases, discrimination and injustice; allowing them to practice compassion, and work towards establishing general welfare to the public.

5 DISCUSSION AND REFLECTION

As our study finding above show, our design fictions evoked a varied range of reaction among both groups of participants. These reactions helped elucidate both participants' values expectations and their perceptions about the technical functioning of those systems, as well as connections across those two concerns.

In the following, we first discuss the inclusion of domain experts in values elicitation process for automated language technologies [59, 86, 118], thereby expanding Value Sensitive Design (VSD) for automated tools. Second, we provide a detailed comparison of our domain experts' automated NLP perceptions with that of technical experts [35, 134–137]. Third, we compare and situate our empirical findings with related prior work in different domains and contexts, including studies that applied design fiction [7, 14, 38, 49, 89, 102, 142]. We conclude by providing critical reflection on our methodological and design decisions, including detailing on our positionality. Throughout, we discuss and identify missed opportunities, suggesting future research and design avenues [115].

5.1 Inclusion of Domain Experts' Values in Automated Language Technology Design

While some efforts has been made to engage industry participants in ethical discussion [e.g., 7, 142], values elicitation around automated computational tools (e.g., ML, NLP) is still scarce. This paper demonstrates an example of engaging domain experts—in this case, journalists and legal experts—in the values elicitation process, as well as the resultant insights gained about specific automated NLP technologies.

Our results indicate that design fictions informed by organizations' data work practices, including challenges and needs, presents an effective way of engaging domains experts in values conversation and in values discovery [86]. Although we designed the speculative concepts around both organizations' values (e.g., transparency, compassion), we did not explicitly describe those values to participants so as to mitigate researcher influence [16, 75]. The values that surfaced through the expression of users' ethical concerns often involved stakeholders who will use or otherwise be indirectly impacted by potential technologies. Therefore, even though participants who are going to use the system take precedence [86], such stakeholders also played a major role in determining values discovery [7]. Thus, the source of these values were not only the design concepts themselves but also the people who are directly and indirectly impacted by those designs, collectively bringing forth our participants' values (e.g., assemblage [118]). Future work should also engage directly with such impacted participants [cf. 141], since *"value decisions must come from domain experts and affected populations"* [57, p. 139, emphasis added].

Our study findings also highlight how participants' values align with those proposed in prior VSD work, including autonomy, invasion of privacy, trust, freedom from biases, and human welfare [58, 60–62]. This alignment increases confidence in the empirical findings for these domains [86]. Put differently, our design fictions enabled participants to bring their tacit value choices from their unconscious and made them available for the researchers to consider them in the future tool design [115]. While some values may be specific to the two non-profits or the two professional domains involved in this work (e.g., justice, transparency), others are likely to be more broadly applicable (e.g., trust). Both domain users put great emphasis on telling the truth to the audience and on adhering to justice [81, 109, 117]. Likewise, the values surfaced pertained to participants' domain expertise and practices that cannot be understood by studying other end user groups [57]. For example, our domain experts engaged in serious conversations, reflecting upon real world consequences of NLP automation based on their years of practical domain knowledge, whereas engaging other end user groups might garner a different reaction, perhaps interesting and useful but based on assumptions, imagination, rather than experiential knowledge. With this study, we set up the stage for the inclusion of values that domain experts desire in future technology design and strengthen their voice for designing technology with them [24, 110]. That said, future work should attend to the interplay between such values and the design of automated systems (NLP or otherwise). For instance, the ways that concern for human autonomy manifests in reactions to DocSelfSorter or to VeriFact may differ from the the ways that same value of autonomy manifests for other domain experts (healthcare practitioners, public service caseworker, etc.) in high-stake algorithmic decision making domain (e.g., public, private services [41, 56, 110, 112]) where human discretion and intelligence is required.

5.2 Automated Tool Perceptions: Technical Experts versus Domain Experts

Despite fundamental differences in professional practices, both data scientists and domain experts work with data and engage in data analysis, either fully manually or with the help of computational tools. Some work has emphasized the similarities between expert data scientists' data work practices and those of journalists [78, 108, 121]. Furthermore, Muller et al. [97] wrote that it is possible that different professionals are doing similar data work. The analysis in this paper reveals both overarching similarities and notable difference when comparing our domain experts' desire with that of expert data scientists' desire from automated NLP systems (e.g., AutoAI/AutoML text tools).

First, similar to data scientists [136], journalists and legal experts did not want black-box automated tools. Rather, they asked for explainable automated NLP tools that shows how it produced a certain outcome, in addition to confidence levels. They also wanted the system to reveal training dataset information (e.g., [120]). A lack of such details made them distrust the system designs, similar to data scientists' lack of confidence in AutoDS [133]. Interestingly, prior work found that domain experts working with data scientists tended to prefer a slightly higher level of automation, whereas our participants' desire for human-guided automation was more comparable to that of expert data scientists [134].

Additionally, journalists and legal experts might also have different automation needs (e.g., full automation in one phase, human-in-the-loop in another) depending on the data analysis phase (e.g., exploration, verification [121]) similar to that of a data scientist [134]. Furthermore, the importance of understanding context associated with data [51, 143], as well as the need for input and validation from external domain experts (such as scientists), to understand the relationships in the data (and across multiple datasets) was important for both our participants as well as data scientists. Due to inherent ambiguity, unreliability, biases, and the need for human discretion [2, 56, 74, 75, 100, 122], such aspects likely cannot be replaced by automated systems.

In terms of contrasts, journalists and legal experts perceived the automated NLP tools as capable of enhancing their current abilities [143], but they also saw the automated NLP tools as a threat to their job security. Data scientists, however, were somewhat less concerned about their jobs being fully automated [133]. Instead, they perceived AutoAI augmenting their work (e.g., enabling them to do more experiments, improve their own code and skills). At the same time, they were also concerned about AutoAI reducing the technical depth or even weakening their technical skills to carry out good data science [136].

Furthermore, data scientists saw AutoML as significantly reducing the amount of time it takes to go from "data to insights" [136], while our participants did not perceive such benefits. Perhaps this difference was more indicative of the fast-paced nature of data science or of profit-oriented [105], time pressured organization context compared to non-profit organizations in this study. Alternatively, perhaps this difference reflects the importance of values for our participants [117, 119].

5.3 Situating Research Findings With Relevant Prior Work

In addition to data scientists, our findings showed important similarities and differences with related prior studies about algorithmic systems in various professional contexts. These findings were comparable regardless of the chosen technological domain, methodology applied, including study goal.

Several of these works were situated in profit-oriented organizations, as opposed to our work in non-profit context where workers work in resource-constrained circumstances (e.g., [23, 91, 93, 130]). For instance, similar to Christin [38], our findings indicates that, despite working under several constraints (e.g., cost, labor, time), our participants did not appreciate full-fledged automated systems. They would rather rely on themselves and adhere to their own judgment instead of blindly following automated system output. They also openly criticized the designs (e.g., open critique [38]), went as far as by saying that they would ignore (e.g., foot-dragging [38]) the tools output in situations if it contradicts with their decisions, because changing their own work based on algorithm prediction could be problematic [56]. While participants were less concerned from their own mishandling of these tools, they showed ethical concerns about other stakeholders' (e.g., defendant, judge) unintended use of these tools (e.g., gaming or manipulation [38]). In contrast with Christin [38] work, she found that most journalists rely on web-analytics for measuring the news impact, and they attached an emotional connotation to it (e.g., pride, shame). Perhaps this difference is due to the study's focus on web analytic tools in profit-oriented newsrooms, rather than automated NLP tools in the non-profit context, where the meaning of impact can differ in significant ways (e.g., enact change, taking risks) [42, 81]. On the contrary, our legal experts reaction were very similar to the lawyers' response in [38], probably, the focus was on risk assessment tool and strong identity (difficulty and formal training required to practice law) made them less receptive of these tools suggestions.

Consistent with Bender et al. [14], who cautioned against when developing tools such as BERT [46], GPT [30] that uses large NLP models, was in alignment with our participants (ethical) concerns of potential risks and biases associated with NLP automation, expressed as their current practices might be able to navigate better [115] using intuition, judgement and subjectivity.

The significance of subject matter expertise (SME) in the decision making was inline with our research findings across different domains [104, 136]. For example, Houde et al. [72], who created dystopian Generative AI futures, and Diakopoulos and Johnson [49] who studied ethical implications to Deepfakes emphasised the importance of a human expert (e.g., moderator) to reduce (e.g., through regulation) the harmful impact of these advanced automated technologies. Unlike some of these study findings, our participants did not recommend any technical intervention (e.g., data traceability and auditability, watermarking [72]; automated verification [49]) against automated technology harm, perhaps due to their background unrelated to technical domains.

When compared to Wu et al. [142], who studied automated tools in the industry profit-oriented context, we faced no challenges to encourage participants to participate in a design fiction workshop, they were excited to take part in the workshop. Similar to Wu et al. [142] and Bardzell et al. [9], we would encourage researchers to exhibit caution when applying conflict as a provocative element, because this could destabilize the research agenda. For example, participant's fear of harm could introduce mistrust [60] in the designs. While our participants were excited to co-speculate with us, at the same time, they were also skeptical about the designs, which was reflected as doubt, and being slightly defensive as both groups of participants practice a great deal of independence in their work, perhaps, due to their focus on individual strong identity and agency. Informed by prior work [76], we clarified before and during the study about the speculative nature of the designs. However, participants' response was to express their desires – by indicating what a design missed and what a design should not do. Again, this workshop being the second encounter with both groups made the critical design interaction more readily accepted than it would have been otherwise. Our research became more participant-centered, changing our initial research inquiry to discovering their values [86], similar to critical design studies where the initial research objective changes and researchers usually become less relevant like the "death of the author" [9]. Evidently, a large body of literature has examined users' ethical concerns successfully [49, 72, 142] in different technology context. However, our study makes a unique contribution that builds upon and expands on prior design fiction work by discovering domain experts' (manageable and non-redundant list) values expectations in the context of automated NLP tools through empirical inquiry [36].

Several prior studies have also examined values elicitation, reflection, and exploration with various users in different technology settings [80, 86, 102, 118, 139]. For example, unlike our study, Wong et al. [139] engaged students in (privacy related) values reflection in the context of biosensing technologies was interested in understanding the designs through values lens and not discovering concrete values per se. In a values exploration study, Le Dantec et al. [86] discovered unique values of the homeless people using mobile technology such as *staying connected*, *identity control*, and tensions with *independence*. Similarly, Odom [102] discovered *resourcefulness* as values for small-scale food producers in the context of interactive systems for sustainable local food production. While most of these values are unique or informed by other frameworks (e.g., Meta-Inventory of Human Values (MIHV) [36, 80]), our domain experts' values expectations were somewhat more aligned with those of values prescribed in the VSD framework [59]. We did not find any new or alternative values [63]. Similar to Odom [102], who showed values tensions with technology augmentation, our domain experts displayed values conflicts with NLP automation. Domain experts also showed conflicting values of *subjectivity and autonomy*, similar to tensions with *independence* in Le Dantec et al. [86] study. These comparisons again illustrate both similarities and differences between our findings and prior work.

5.4 Researcher Reflection

In this section, we will present our reflection on each of our design decisions, borrowing reflective constructs from [84, 115] as a guide. Our reflection is to suggest route that design futuring researchers might take, or decide to avoid in their work.

5.4.1 Researcher Positionality: Why Automated NLP Future? According to Kozubaev et al. [84], design futures might 'close down' and 'opens up' other possible, preferable, or problematic futures. Specifically, in our case, after an initial literature review for inspiration and ideas both in design futuring work and current NLP technology, we envisaged our automated NLP systems futures informed by the challenges participants described during the interviews. The challenges were translated into various automated NLP tools to support them at various phases of their data work practices. We then organized our design concepts around their data work practices: exploration, verification, and publication. This is why the design fictions were not co-produced with them, because our main objective was to explicate their perceptions about automated text tools [119] before developing any technology for them.

We designed these concepts for the future inspired by NLP technologies that are possible now. Hence, futures were informed by users recent-past data analysis actions and challenges. Time was also used as a resource; in other words, manual text processing took a long time, thus, our automated fictional NLP tools tried to minimize the overhaul by doing things for them. Automated NLP futures played two key purposes in our designerly quest: first, it was applied to ease in participant's manual text data processing; second, it acted as a provocative element. Provocativeness can be implemented by adding conflicts in the design [9]. To balance unforeseen friction, we incorporated other indirect users perspective into the designs to indicate that the designs were not just about them but it might also impact (e.g., benefit or harm) other users (e.g., victims, perpetrator) in different ways. In the process, we created NLP futures that were free from explicit dystopian experiences, (unlike Houde et al. [72]). We intentionally left those harmful aspects to invoke participant's thought processes. The automation aspect of NLP was a successful conflict element that evoked a range of reactions (both individual and collaborative [96]). In that way, we used design fiction to co-speculate with the participants (implication, ideation), to learn about their values expectations, neither the underlying details of model workings [144], nor solutions [20, 21].

5.4.2 Why Scrapbook of Provocations? Our design falls into the category of critical design representation of provocations, while shares a lot of similarity of construction with that of designer's workbook style.

Our design fiction scrapbook sits somewhere at a hybrid space constructed by reusing several concepts drawing from fictional literature of workbooks and provocations (see section 3.3.1 and 3.3.2 for more details). For instance, provocativeness was used to introduce conflict by removing humans [142], textual descriptions were intentionally left incomplete to encourage open interpretation [116] and ambiguity of information [66]. The visual elements were used to provide context within which the designs lived in an imagined world [140] including people (various direct/indirect/excluded stakeholders) and their relationships with the technology. For example, a participant imagined himself at the airport wearing DataWatch, he pushed the design by saying that lawyars don’t work that way (e.g., shows ambiguity of context [66]), some even described how annoying would that be to have watch notification during a personal meeting (inhibiting that world [40]). Thus, our designs along with these contexts took them in an imaginary world into the future, while staying in the present [53]. Therefore, we took previous design inspiration to build a new world, to address a different problem with a two different domain expert users. In that way, we expand on previous design futuring work and suggest that we can build off existing fictional work and effectively strategize them to address different issues and contexts.

Another important difference of our proposal with existing fictional designs such as cultural probe [64], workbook [139, 140], and designers booklets [43] was physicality in terms of whether it is printed or not, something that people could carry around or it required mandatory return or not. Compared to Harrington and Dillahunt [70] workbook which required mandatory filling out information and shared them via Google Drive, whereas our scrapbook required no mandatory filling out of information. We suspect that if participants were asked to return a version of the fictional object (e.g., probe such as post cards, maps [64, 86]), that might evoke a different reaction.

Rather than inviting participants to write stories or scenarios [3, 49, 72], we presented participants with a series of design concepts and asked them to imagine inhabiting the world where those concepts exist. Our participants response indicate that our design fictions enabled them to reflect on the designs as well as their own work practices thoroughly and thoughtfully [115], probably, emailing the scrapbook a week ahead before the workshop was useful to afford such reaction. Another important point of contrast was using initial exploratory interviews to inform the design fictions and situate them in data work phases (see section 3.2.1), rather than just searching for cues in the interviews [142]. This method helped us to have a shared vocabulary and exchange among participants, researchers, and the designs. [114].

6 LIMITATIONS AND FUTURE WORK

There are several ways, this study can be extended in the future. First, the values identified from this study represents values that our participants care about afforded by our design fictions and the fictional world in which they were situated. It is likely that these values may not be representative of journalists/legal experts working in different organizations (e.g., across different race, nationality [39, 145]). Second, future work might further expand one/more designs, for example, how human-guided automation would look like for DocSelfSorter and conduct further studies. This might generate a different discussion, might even discover new or alternate values, and other societal and ethical implications. Finally, domain experts from diverse backgrounds could be engaged to offer value expectations in establishing a theoretical framework of design guidelines for automated NLP (or AI automation more broadly) for domain experts.

7 CONCLUSION

In this study, we discussed findings from two participatory design fiction workshops where we engaged domain experts who are not professionally trained in computational tools/techniques to discuss their values expectations from several speculative automated text processing systems. Overall, our study makes two contributions, first, our case study shows that it is possible to create distinct futures for specific domains by building off of the previous design fiction work. This new world effectively engaged users in values conversations. Second, we also report on our empirical findings about the values our domain experts care about, including their perception that further shed lights on tensions with automation and implications for how not-to-design them. Lastly, our findings provides the groundwork for engaging domain experts whose expertise lies outside of the field of computation in values elicitation for automated systems. Such users often excluded from important technology ethical conversations. With this research, we have just scratched the surface and suggest researchers to further examine this under explored avenue.

ACKNOWLEDGMENTS

We would like to thank all of our study participants for their time and dedication. We are thankful to the anonymous reviewers and AC’s for valuable comments and feedback. This material is based on work supported in part by the NSF under Grant No. IIS-1844901.

REFERENCES

[1] Demi Ajayi. 2020. What to consider when using BERT and GPT models for NLP. https://www.ibm.com/blogs/watson/2020/12/what-to-consider-when-using-bert-and-gpt-models-for-nlp/ Accessed: 2021-03-28.

[2] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[3] Aloha Hufana Ambe, Margot Brereton, Alessandro Soro, Laurie Buys, and Paul Roe. 2019. The adventures of older authors: Exploring futures through co-design fictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[5] Jodi Aronson. 1995. A pragmatic view of thematic analysis. *The qualitative report* 2, 1 (1995), 1–3.

[6] Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. 2016. Opinion mining and sentiment analysis. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 452–455.

[7] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 421–433.

[8] Jeffrey Bardzell and Shaowen Bardzell. 2013. What is" critical" about critical design?. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3297–3306.

[9] Shaowen Bardzell, Jeffrey Bardzell, Jodi Forlizzi, John Zimmerman, and John Antanitis. 2012. Critical design and critical theory: the challenge of designing for provocation. In *Proceedings of the Designing Interactive Systems Conference*. 288–297.

[10] Eric P. S. Baumer and Jed R. Brubaker. 2017. Post-Userism. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Denver, CO, 6291–6303. https://doi.org/10.1145/3025453.3025740

[11] Eric P. S. Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1472–1482.

[12] Eric P. S. Baumer and M Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2271–2274.

[13] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[14] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[15] Carlo Biagioli, Enrico Francesconi, Andrea Passerini, Simonetta Montemagni, and Claudia Soria. 2005. Automatic semantics extraction in law documents. In *Proceedings of the 10th international conference on Artificial intelligence and law*. 133–140.

[16] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The Values Encoded in Machine Learning Research. arXiv:2106.15590 [cs.LG]

[17] Ekaba Bisong. 2019. Google AutoML: Cloud Natural Language Processing. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 599–612.

[18] Julian Bleecker. 2009. Design fiction. A short essay on design, science, fact and fiction. Near Future Laboratory (2009).

[19] Mark Blythe. 2017. Research fiction: storytelling, plot and design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5400–5411.

[20] Mark Blythe, Kristina Andersen, Rachel Clarke, and Peter Wright. 2016. Anti-solutionist strategies: Seriously silly design fiction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4968–4978.

[21] Mark Blythe and Enrique Encinas. 2018. Research Fiction and Thought Experiments in Design. *Foundations and Trends® Human–Computer Interaction* 12, 1 (May 2018), 1–105. https://doi.org/10.1561/1100000070

[22] Mark Blythe, Enrique Encinas, Jofish Kaye, Miriam Lueck Avery, Rob McCabe, and Kristina Andersen. 2018. Imaginary design workbooks: Constructive criticism and practical provocation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[23] Chris Bopp, Ellie Harmon, and Amy Voida. 2017. Disempowered by data: Nonprofits, social enterprises, and the consequences of data-driven work. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3608–3619.

[24] Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1125–1134.

[25] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[26] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic analysis. *Handbook of Research Methods in Health Social Sciences* (2019), 843–860.

[27] Sarah Brayne and Angèle Christin. [n.d.]. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems* ([n. d.]). https://doi.org/10.1093/socpro/spaa004

[28] Meredith Broussard. 2015. Artificial intelligence for investigative reporting: Using an expert system to enhance journalists' ability to discover original public affairs stories. *Digital Journalism* 3, 6 (2015), 814–831.

[29] Barry Brown, Julian Bleecker, Marco D'adamo, Pedro Ferreira, Joakim Formo, Mareike Glöss, Maria Holm, Kristina Höök, Eva-Carin Banka Johnson, Emil Kaburuan, et al. 2016. The IKEA Catalogue: Design fiction in academic and industrial collaborations. In *Proceedings of the 19th International Conference on Supporting Group Work*. 335–344.

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

[31] Bill Buxton. 2007. *Sketching User Experiences: Getting the Design Right and Getting the Right Design*. Morgan Kaufmann, San Francisco, CA.

[32] Stuart Candy and Jake Dunagan. 2017. Designing an experiential scenario: The people who vanished. *Futures* 86 (2017), 136–153.

[33] David Caswell and Konstantin Dörr. 2018. Automated Journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism practice* 12, 4 (2018), 477–496.

[34] Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

[35] Arunima Chaudhary, Alayt Issak, Kiran Kate, Yannis Katsis, Abel Valente, Dakuo Wang, Alexandre Evfimievski, Sairam Gurajada, Ban Kawas, Cristiano Malossi, et al. 2021. AutoText: An End-to-End AutoAI Framework for Text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 16001–16003.

[36] An-Shou Cheng and Kenneth R Fleischmann. 2010. Developing a meta-inventory of human values. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[37] Fanny Chevalier, Melanie Tory, Bongshin Lee, Jarke van Wijk, Giuseppe Santucci, Marian Dörk, and Jessica Hullman. 2018. From analysis to communication: Supporting the lifecycle of a story. In *Data-Driven Storytelling*. AK Peters/CRC Press, 151–183.

[38] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.

[39] Sasha Costanza-Chock. 2018. Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society* (2018).

[40] Paul Coulton, Joseph Galen Lindley, Miriam Sturdee, and Michael Stead. 2017. Design Fiction as World Building. In *Proceedings of Research through Design Conference (RTD)*. Edinburgh, Soctland.

[41] Kate Crawford. 2021. The atlas of AI. In *The Atlas of AI*. Yale University Press.

[42] Hugo De Burgh. 2008. *Investigative journalism*. Routledge.

[43] Audrey Desjardins, Cayla Key, Heidi R Biggs, and Kelsey Aschenbeck. 2019. Bespoke booklets: A method for situated co-speculation. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 697–709.

[44] DoNotPay Developers. February 2021. DoNotPay. https://donotpay.com/ Accessed: 2021-02-18.

[45] Temi Developers. November 2019. Advanced speech recognition software. https://www.temi.com Accessed: 2019-11-12.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[47] Nicholas Diakopoulos. 2018. Ethics in Data-Driven Visual Storytelling. In *Data-Driven Storytelling*. AK Peters/CRC Press, 233–248.

[48] Nicholas Diakopoulos. 2019. *Automating the news*. Harvard University Press.

[49] Nicholas Diakopoulos and Deborah Johnson. 2019. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society* (2019), 1461444820925811.

[50] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic transparency in the news media. *Digital journalism* 5, 7 (2017), 809–828.

[51] Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism*. MIT Press.

[52] Anthony Dunne and Fiona Raby. 2001. *Design noir: The secret life of electronic objects*. Springer Science & Business Media.

[53] Anthony Dunne and Fiona Raby. 2013. *Speculative everything: design, fiction, and social dreaming*. MIT press.

[54] Tony Dunne and Fiona Raby. 2001. *Design Noir: The Secret Life of Electronic Objects*. Birkhäuser, Berlin.

[55] Pelle Ehn. 1988. *Work-oriented design of computer artifacts*. Ph.D. Dissertation. Arbetslivscentrum.

[56] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[57] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[58] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[59] Batya Friedman, Peter H Kahn, and Alan Borning. 2008. Value sensitive design and information systems. *The handbook of information and computer ethics* (2008), 69–101.

[60] Batya Friedman, Peter H Khan Jr, and Daniel C Howe. 2000. Trust online. *Commun. ACM* 43, 12 (2000), 34–40.

[61] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[62] Batya Friedman and Helen Nissenbaum. 1997. Software agents and user autonomy. In *Proceedings of the first international conference on Autonomous agents*. 466–469.

[63] Bill Gaver and Heather Martin. 2000. Alternatives: exploring information appliances through conceptual design proposals. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 209–216.

[64] William Gaver. 2007. Cultural commentators: Non-native interpretations as resources for polyphonic assessment. *International journal of human-computer studies* 65, 4 (2007), 292–305.

[65] William Gaver. 2011. Making spaces: how design workbooks work. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1551–1560.

[66] William W Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 233–240.

[67] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Daumé, Hal, III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. https://doi.org/10.1145/3458723

[68] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1968. The discovery of grounded theory; strategies for qualitative research. *Nursing research* 17, 4 (1968), 364.

[69] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. 2012. *The data journalism handbook: how journalists can use data to improve the news.* " O’Reilly Media, Inc.".

[70] Christina Harrington and Tawanna R Dillahunt. 2021. Eliciting Tech Futures Among Black Young Adults: A Case Study of Remote Speculative Co-Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–15.

[71] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1803–1812.

[72] Stephanie Houde, Vera Liao, Jacquelyn Martino, Michael Muller, David Piorkowski, John Richards, Justin Weisz, and Yunfeng Zhang. 2020. Business (Mis)Use Cases of Generative AI. In *Human-AI Co-Creation with Generative Models, Workshop at ACM Conference in Intelligent User Interfaces.* 6.

[73] Aniket Jain, Bhavya Sharma, Paridhi Choudhary, Rohan Sangave, and William Yang. 2018. Data-Driven Investigative Journalism For Connectas Dataset. *arXiv preprint arXiv:1804.08675* (2018).

[74] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed: 2020-07-08.

[75] Pratyusha Kalluri. 2020. Don’t ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020), 169–169.

[76] Vera Khovanskaya, Eric P. S. Baumer, and Phoebe Sengers. 2015. Double Binds and Double Blinds: Evaluation Tactics in Critically Oriented HCI. In *Decennial Aarhus Conference on Human Centered Computing*, Vol. 1. 12. https://doi.org/10.7146/aahcc.v1i1.21266

[77] David Kirby. 2010. The Future Is Now: Diegetic Prototypes and the Role of Popular Films in Generating Real-World Technological Development. *Social Studies of Science* 40, 1 (Feb. 2010), 41–70. https://doi.org/10.1177/0306312709338325

[78] Keith Kirkpatrick. 2015. Putting the data science into journalism.

[79] Cory Knobel and Geoffrey C Bowker. 2011. Values in design. *Commun. ACM* 54, 7 (2011), 26–28.

[80] Jes A Koepfler, Katie Shilton, and Kenneth R Fleischmann. 2013. A stake in the issue of homelessness: Identifying values of interest for design in online communities. In *Proceedings of the 6th International Conference on Communities and Technologies.* 36–45.

[81] Bill Kovach and Tom Rosenstiel. 2014. *The elements of journalism: What newspeople should know and the public should expect.* Three Rivers Press (CA).

[82] Sandjar Kozubaev and Carl DiSalvo. [n.d.]. Participatory Workbooks For Speculation: Exploring Library Futures. ([n. d.]).

[83] Sandjar Kozubaev and Carl DiSalvo. 2020. The Future of Public Libraries as Convivial Spaces: A Design Fiction. In *Companion of the 2020 ACM International Conference on Supporting Group Work.* 83–90.

[84] Sandjar Kozubaev, Chris Elsden, Noura Howell, Marie Louise Juul Søndergaard, Nick Merrill, Britta Schulte, and Richmond Y Wong. 2020. Expanding Modes of Reflection in Design Futuring. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–15.

[85] Hoang Thanh Lam, Beat Buesser, Hong Min, Tran Ngoc Minh, Martin Wistuba, Udayan Khurana, Gregory Bramble, Theodoros Salonidis, Dakuo Wang, and Horst Samulowitz. 2021. Automated Data Science for Relational Data. In *2021 IEEE 37th International Conference on Data Engineering (ICDE).* IEEE, 2689–2692.

[86] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. 2009. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 1141–1150.

[87] Carl-Gustav Lindén, Hanna Tuulonen, Asta Bäck, Nicholas Diakopoulos, Mark Granroth-Wilding, Lauri Haapanen, Leo Leppänen, Magnus Melin, Tom Moring, Myriam Munezero, et al. 2019. News automation: The rewards, risks and realities of ‘machine journalism’. (2019).

[88] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding.* 48–55.

[89] Sus Lyckvi, Virpi Roto, Elizabeth Buie, and Yiying Wu. 2018. The role of design fiction in participatory design processes. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction.* 976–979.

[90] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.

[91] Kelly McBride and Tom Rosenstiel. 2013. *The new ethics of journalism: Principles for the 21st century.* CQ Press.

[92] Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807* (2020).

[93] Cecelia Merkel, Umer Farooq, Lu Xiao, Craig Ganoe, Mary Beth Rosson, and John M Carroll. 2007. Managing technology use and learning in nonprofit community organizations: methodological challenges and opportunities. In *Proceedings of the 2007 symposium on Computer human interaction for the management of information technology.* 8–es.

[94] Paola Merlo, James Henderson, Gerold Schneider, and Eric Wehrli. 2003. Learning document similarity using natural language processing. (2003).

[95] Wai Yin Mok and Jonathan R Mok. 2019. Legal Machine-Learning Analysis: First Steps towards AI Assisted Legal Research. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law.* 266–267.

[96] Michael Muller, Jeffrey Bardzell, EunJeong Cheon, Norman Makoto Su, Eric P. S. Baumer, Casey Fiesler, Ann Light, and Mark Blythe. 2020. Understanding the Past, Present, and Future of Design Fictions. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–8.

[97] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–15.

[98] Michael J Muller. 2009. Participatory design: the third space in HCI. In *Human-computer interaction.* CRC press, 181–202.

[99] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[100] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism.* nyu Press.

[101] nostoros. 2018. CONNECTAS. https://www.connectas.org/sobre-nosotros/ Accessed: 2020-08-10.

[102] William Odom. 2010. " Mate, we don’t need a chip to tell us the soil’s dry" opportunities for designing interactive systems to support urban food production. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems.* 232–235.

[103] Cathy O’neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.

[104] Soya Park, April Yi Wang, Ban Kawas, Q Vera Liao, David Piorkowski, and Marina Danilevsky. 2021. Facilitating knowledge sharing from domain experts to data scientists for building nlp models. In *26th International Conference on Intelligent User Interfaces.* 585–596.

[105] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[106] Brent N Reeves and Frank Shipman. 1996. Tacit knowledge: icebergs in collaborative design. *ACM SIGOIS Bulletin* 17, 3 (1996), 24–33.

[107] Thomson Reuters. 1975. WestLaw. https://legal.thomsonreuters.com/en/products/westlaw Accessed: 2020-26-08.

[108] Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. 2018. *Data-driven storytelling.* CRC Press.

[109] David S Romantz and Kathleen Elliott Vinson. 1998. *Legal analysis: The fundamental skill.* Vol. 2. Carolina Academic Press.

[110] Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The Deskilling of Domain Expertise in AI Development. (2022).

[111] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[112] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the us child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–15.

[113] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[114] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph’Jofish’ Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility.* 49–58.

[115] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph ’Jofish’ Kaye. 2005. Reflective Design. In *Decennial Aarhus Conference on Human Centered Computing.* Association for Computing Machinery, Aarhus, Denmark.

[116] Phoebe Sengers and Bill Gaver. 2006. Staying open to interpretation: engaging multiple meanings in design and evaluation. In *Proceedings of the 6th conference on Designing Interactive systems.* 99–108.

[117] Ivor Shapiro. 2010. Evaluating journalism: Towards an assessment framework for the practice of journalism. *Journalism Practice* 4, 2 (2010), 143–162.

[118] Katie Shilton, Jes A Koepfler, and Kenneth R Fleischmann. 2014. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing.* 426–435.

[119] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[120] Dilruba Showkat. 2022. Supporting Responsible Data and Algorithmic Practices in The News Media. In *Proceedings of the 6th HUMANIZE Workshop.*

[121] Dilruba Showkat and Eric P. S. Baumer. 2021. Where Do Stories Come From? Examining the Exploration Process in Investigative Data Journalism. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 390:1–390:31.

https://doi.org/10.1145/3479534

[122] Spencer Soper. 2021. Fired by Bot at Amazon: 'It's You Against the Machine'. https://www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out?sref=EJ3iffSv Accessed: 2021-06-30.

[123] Bruce Sterling. 2013. Fantasy Prototypes and Real Disruption. https://www.youtube.com/watch?v=M7KErICTSHU.

[124] Lucy Suchman. 2006. *Human-Machine Reconfigurations: Plans and Situated Actions* (second ed.). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511808418

[125] Saiki Tanabe. 2017. Sketchboard. https://sketchboard.io/ Accessed: 2020-03-11.

[126] Joshua Tanenbaum, Marcel Pufal, and Karen Tanenbaum. 2016. The limits of our imagination: design fiction as a strategy for engaging with dystopian futures. In *Proceedings of the Second Workshop on Computing within Limits*. 1–9.

[127] Joshua Tanenbaum, Karen Tanenbaum, and Ron Wakkary. 2012. Steampunk as design fiction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1583–1592.

[128] Steven J Taylor and Robert Bogdan. 1984. *Introduction to qualitative research methods: The search for meanings*. Wiley-Interscience.

[129] Ike Vayansky and Sathish AP Kumar. 2020. A review of topic modeling methods. *Information Systems* (2020), 101582.

[130] Amy Voida, Ellie Harmon, and Ban Al-Ani. 2011. Homebrew databases: Complexities of everyday information management in nonprofit organizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 915–924.

[131] Joseph Voros. 2001. A Primer on Futures Studies, Foresight and the Use of Scenarios. *prospect, the Foresight Bulletin* 6, December (2001).

[132] Ron Wakkary, Doenja Oogjes, Henry WJ Lin, and Sabrina Hauser. 2018. Philosophers living with the tilting bowl. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[133] Dakuo Wang, Josh Andres, Justin D Weisz, Erick Oduor, and Casey Dugan. 2021. AutoDS: Towards Human-Centered Automation of Data Science. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[134] Dakuo Wang, Q Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. 2021. How Much Automation Does a Data Scientist Want? *arXiv preprint arXiv:2101.03970* (2021).

[135] Dakuo Wang, Parikshit Ram, Daniel Karl I Weidele, Sijia Liu, Michael Muller, Justin D Weisz, Abel Valente, Arunima Chaudhary, Dustin Torres, Horst Samulowitz, et al. 2020. AutoAI: Automating the End-to-End AI Lifecycle with Humans-in-the-Loop. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*. 77–78.

[136] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[137] Daniel Karl I Weidele, Justin D Weisz, Erick Oduor, Michael Muller, Josh Andres, Alexander Gray, and Dakuo Wang. 2020. AutoAIViz: opening the blackbox of automated artificial intelligence with conditional parallel coordinates. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 308–312.

[138] Langdon Winner. 2017. *Do artifacts have politics?* Routledge.

[139] Richmond Y Wong, Deirdre K Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. 2017. Eliciting values reflections by engaging privacy futures using design workbooks. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–26.

[140] Richmond Y Wong, Ellen Van Wyk, and James Pierce. 2017. Real-fictional entanglements: Using science fiction and design fiction to interrogate sensing technologies. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 567–579.

[141] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Montréal, QC, 656:1–656:14. https://doi.org/10.1145/3173574.3174230

[142] Yiying Wu, Sus Lyckvi, and Virpi Roto. 2019. " What is Fair Shipping, Anyway?" Using Design Fiction to Raise Ethical Awareness in an Industrial Context. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[143] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. 2021. Whither automl? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[144] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 185.

[145] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

A RESEARCH METHODS

A.1 Background Semi-structured Interview Questions

This portion of the appendix outlines some of the questions asked of participants during the background interviews. Since these were semi-structured interviews, exact questions and follow-up probes were tailored based on participants' responses.

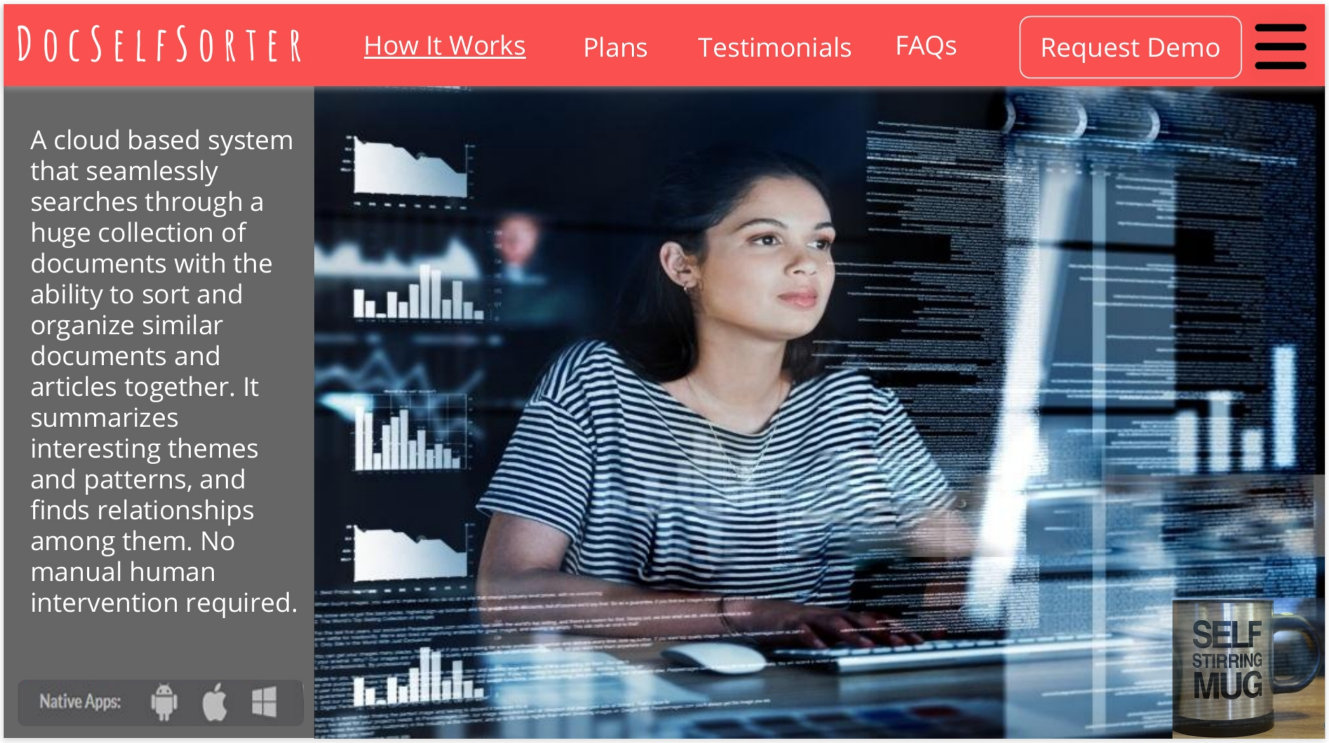
- Personal details: Please tell me about yourself.
- Role: What is your current role at [organization name]? Please tell me about your role?
- Specific Project: Could you please talk about a project that you have completed in the past?
- Data Source: What were the data sources you used in this project? How did you learn about the data sources? How did you get the data? What was your data selection process?
- Data Analysis: What were the tools that you use for data analysis? What was your go-to process for data analysis?
- Data Summarization: What was the process of transferring data to report preparation?
- Sketching: The interviewer provided a prompt along these lines for the sketching activity. "We talked about various data sources, tasks, people, data selection, data analysis tools, data validation, the output, and the story. Now, please sketch how all these activities go together in your reporting task. Include all the data sources, people, and tasks that you considered for the project. Include all the decisions that you had to make. Include all the actions that you had to take."

A.2 Workshop Question Prompts

This portion of the appendix outlines some of the questions asked of participants during the workshop. Exact questions and follow-ups varied, based on whether the workshop was being conducted at JME or PP (examples below are from JME), on the specific design fiction concept being discussed (examples below pertain to DataWatch), and on participants' responses.

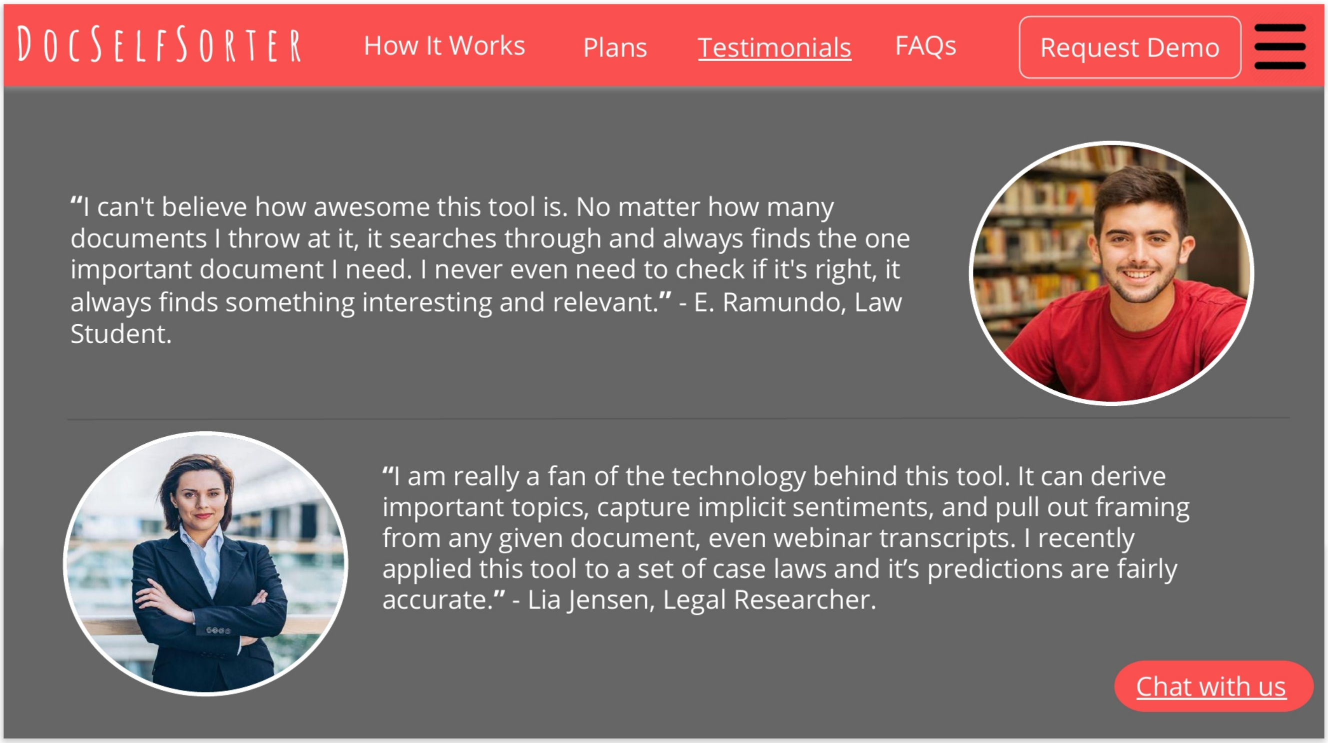
- In just a few words, can you briefly describe your reaction to the DataWatch?
- If you had access to the DataWatch, what are examples of laws or on-going cases would you use it to track?
- If you were to envision yourself using DataWatch, rather than the systems you currently use to track these kinds of legal developments, how would your work be different?
- In what ways does the DataWatch reflect (or possibly not reflect) JME's core values?
- What concerns would you have about a future world where a DataWatch exists?

B DESIGN FICTIONS FOR LEGAL ANALYSIS



A cloud based system that seamlessly searches through a huge collection of documents with the ability to sort and organize similar documents and articles together. It summarizes interesting themes and patterns, and finds relationships among them. No manual human intervention required.

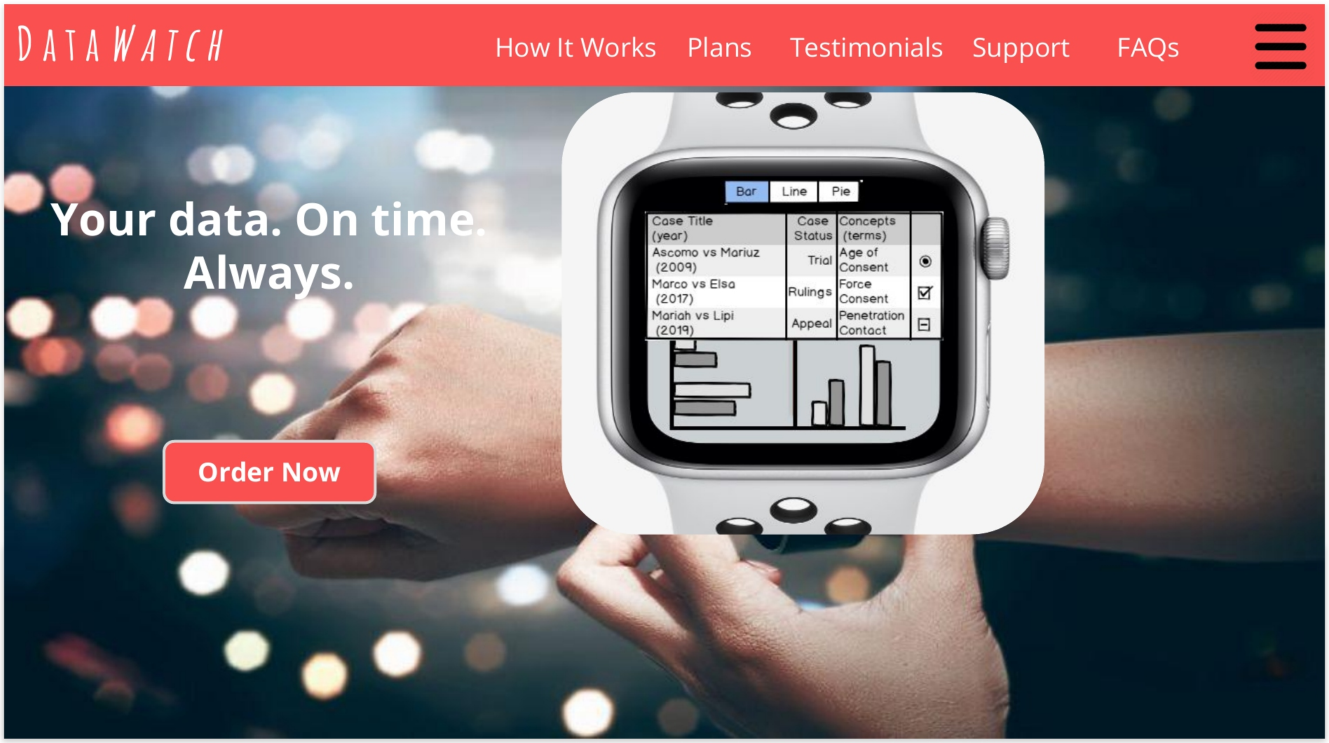
Figure 2: A web page describing how DocSelfSorter works.



"I can't believe how awesome this tool is. No matter how many documents I throw at it, it searches through and always finds the one important document I need. I never even need to check if it's right, it always finds something interesting and relevant." - E. Ramundo, Law Student.

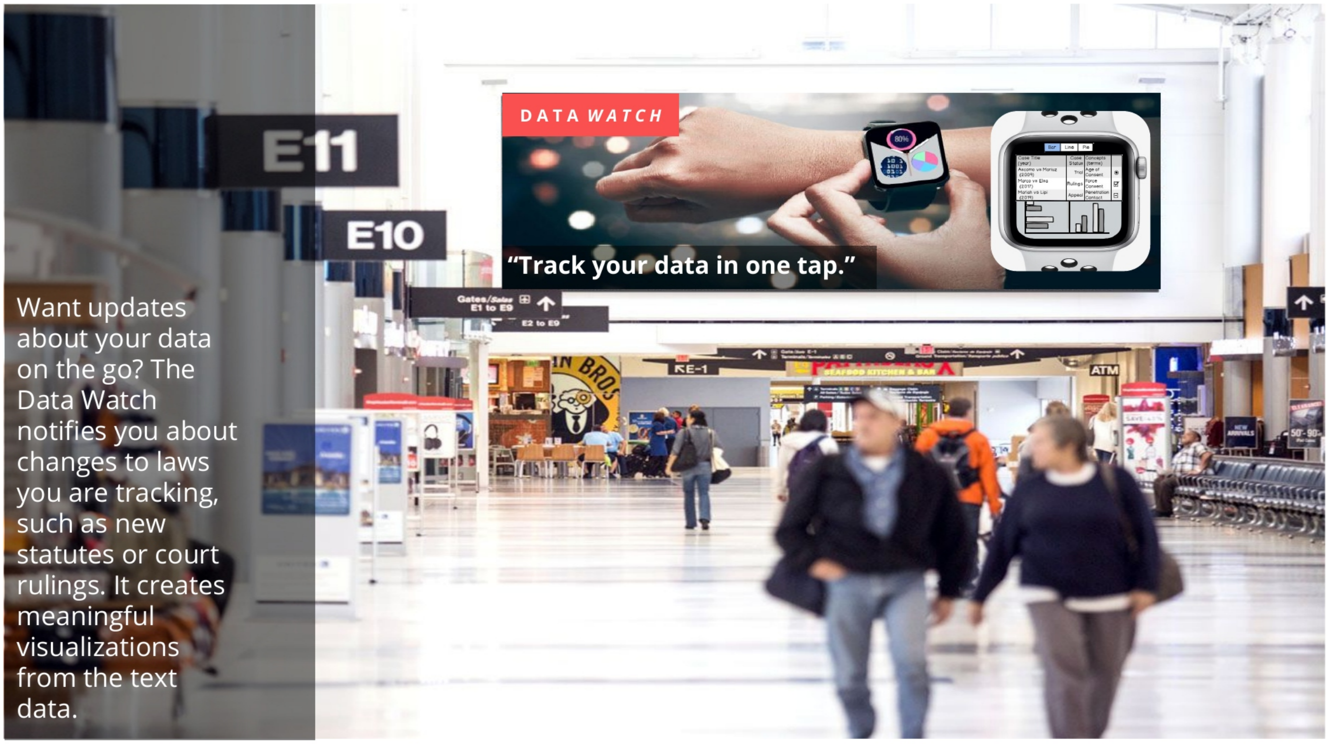
"I am really a fan of the technology behind this tool. It can derive important topics, capture implicit sentiments, and pull out framing from any given document, even webinar transcripts. I recently applied this tool to a set of case laws and it's predictions are fairly accurate." - Lia Jensen, Legal Researcher.

Figure 3: DocSelfSorter presented via user testimonials from a law student and a legal professional perspective.



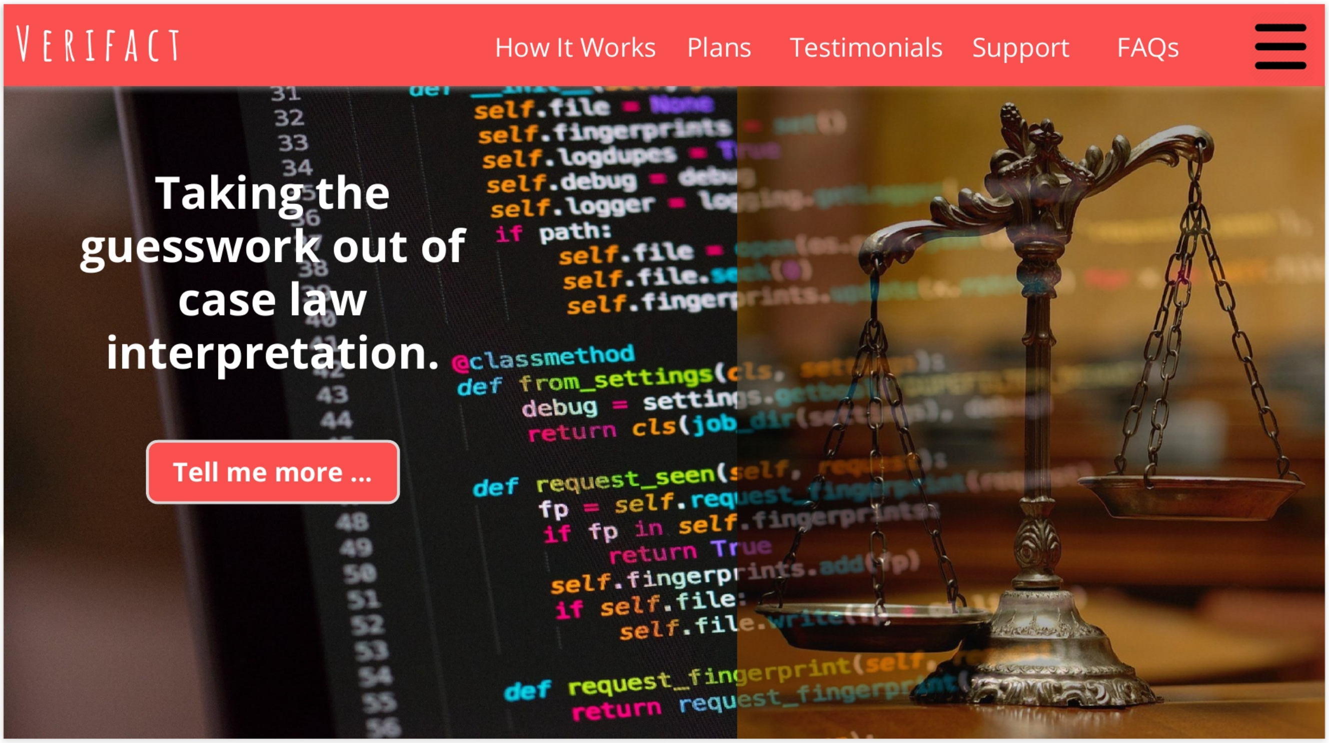
Case Title (year)	Case Status	Concepts (terms)	
Ascomo vs Mariuz (2009)	Trial	Age of Consent	◉
Marco vs Elsa (2017)	Rulings	Force Consent	☑
Mariah vs Lipi (2019)	Appeal	Penetration Contact	⊟

Figure 4: DataWatch presented as a landing page with a tagline indicating the purpose of the design.



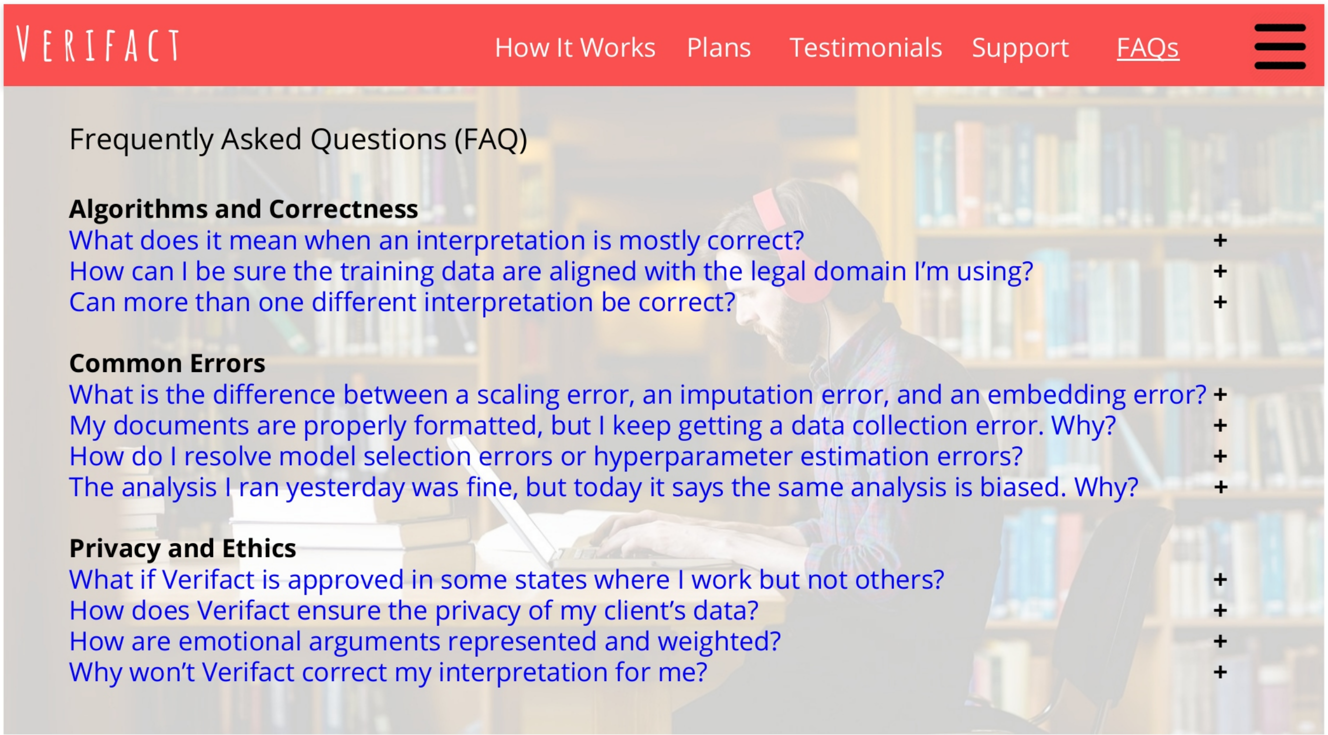
Want updates about your data on the go? The Data Watch notifies you about changes to laws you are tracking, such as new statutes or court rulings. It creates meaningful visualizations from the text data.

Figure 5: DataWatch advertised as an airport billboard add; left paragraph text describing its capabilities.



Taking the guesswork out of case law interpretation.

Figure 6: VeriFact shown as a landing page. Tagline hinting at replacing ambiguity in case law interpretation.



Frequently Asked Questions (FAQ)

Algorithms and Correctness
What does it mean when an interpretation is mostly correct? +
How can I be sure the training data are aligned with the legal domain I'm using? +
Can more than one different interpretation be correct? +

Common Errors
What is the difference between a scaling error, an imputation error, and an embedding error? +
My documents are properly formatted, but I keep getting a data collection error. Why? +
How do I resolve model selection errors or hyperparameter estimation errors? +
The analysis I ran yesterday was fine, but today it says the same analysis is biased. Why? +

Privacy and Ethics
What if Verifact is approved in some states where I work but not others? +
How does Verifact ensure the privacy of my client's data? +
How are emotional arguments represented and weighted? +
Why won't Verifact correct my interpretation for me? +

Figure 7: VeriFact imagined on a webpage showing frequently asked questions (FAQs).

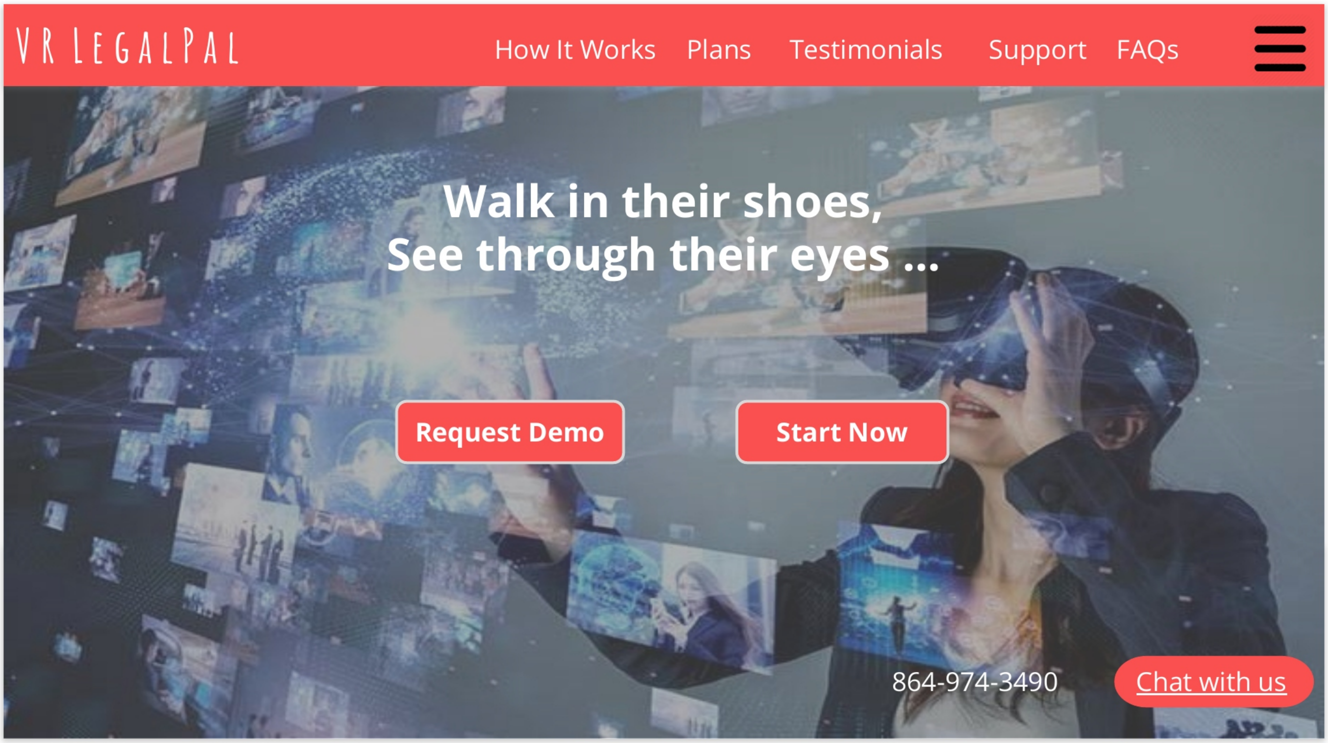
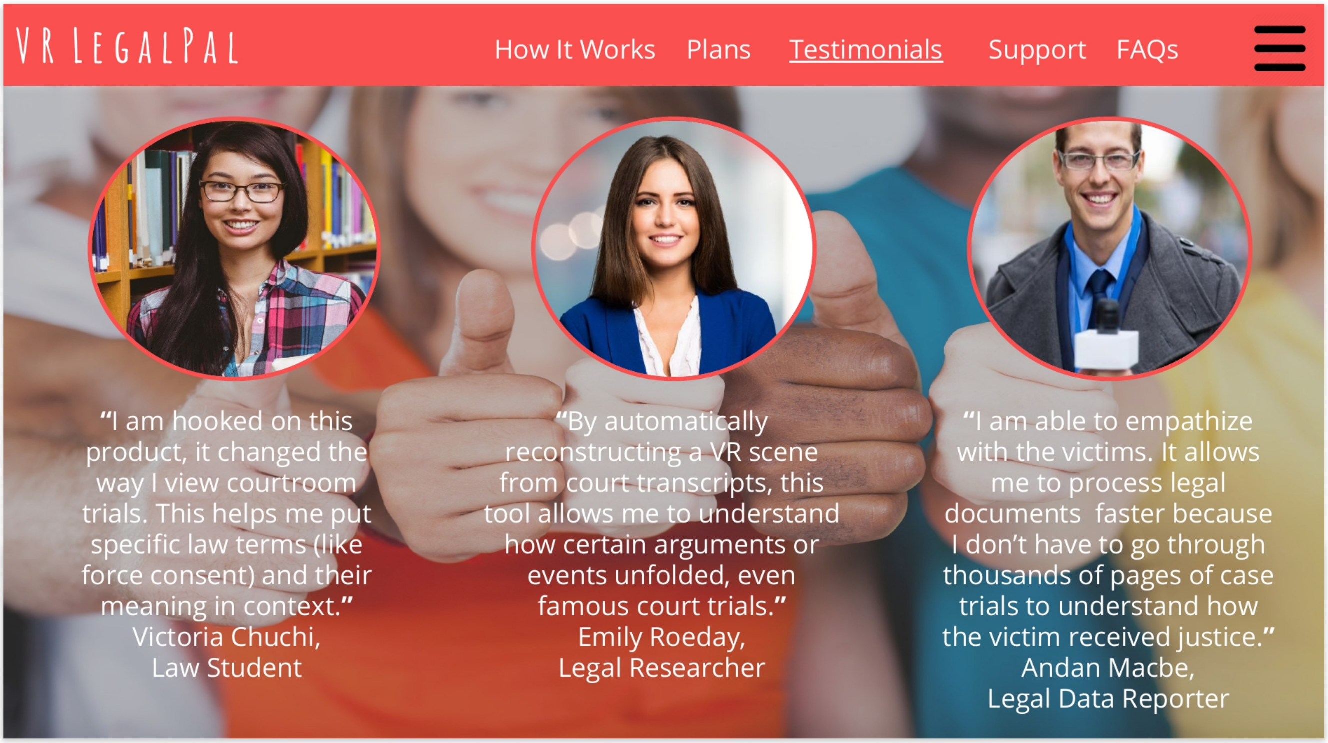


Figure 8: VRLegalPal imagined on a landing page. Tagline emphasizing to take on others' (e.g., victim, perpetrator) perspec-

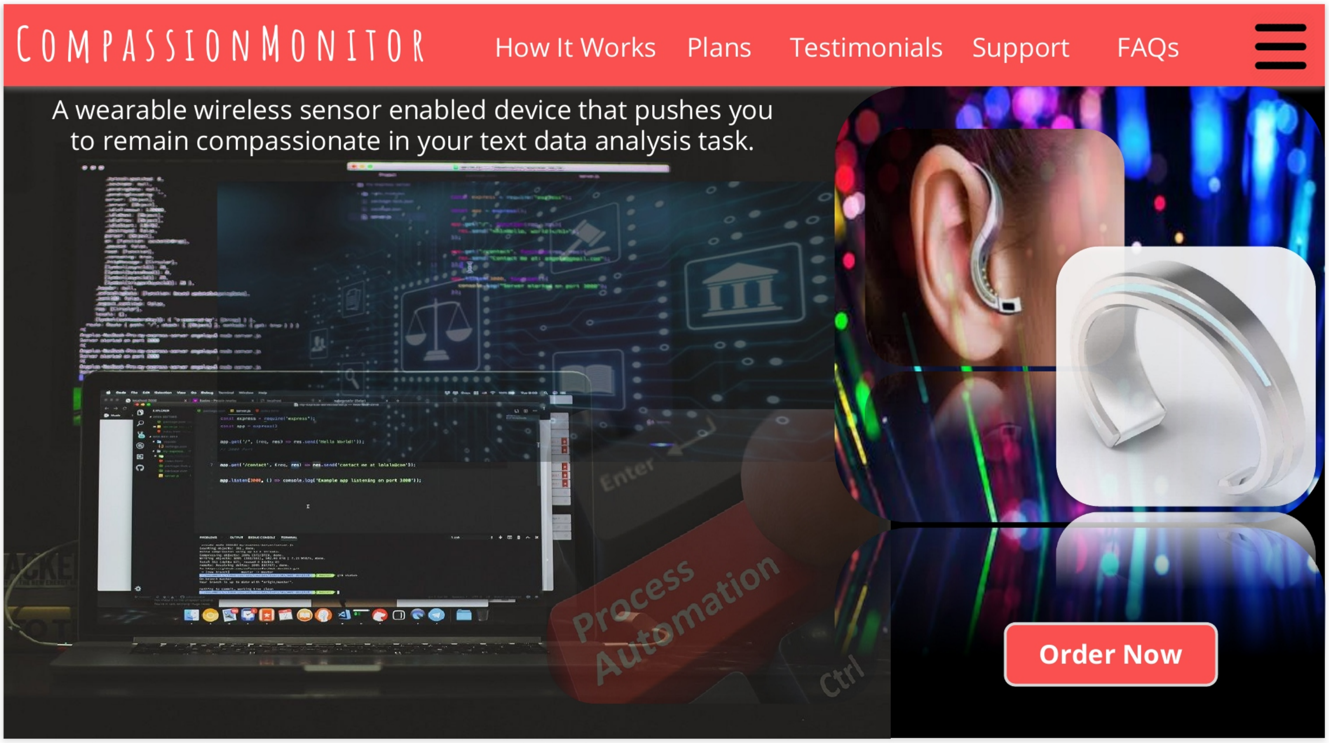


"I am hooked on this product, it changed the way I view courtroom trials. This helps me put specific law terms (like force consent) and their meaning in context."
Victoria Chuchi,
Law Student

"By automatically reconstructing a VR scene from court transcripts, this tool allows me to understand how certain arguments or events unfolded, even famous court trials."
Emily Roeday,
Legal Researcher

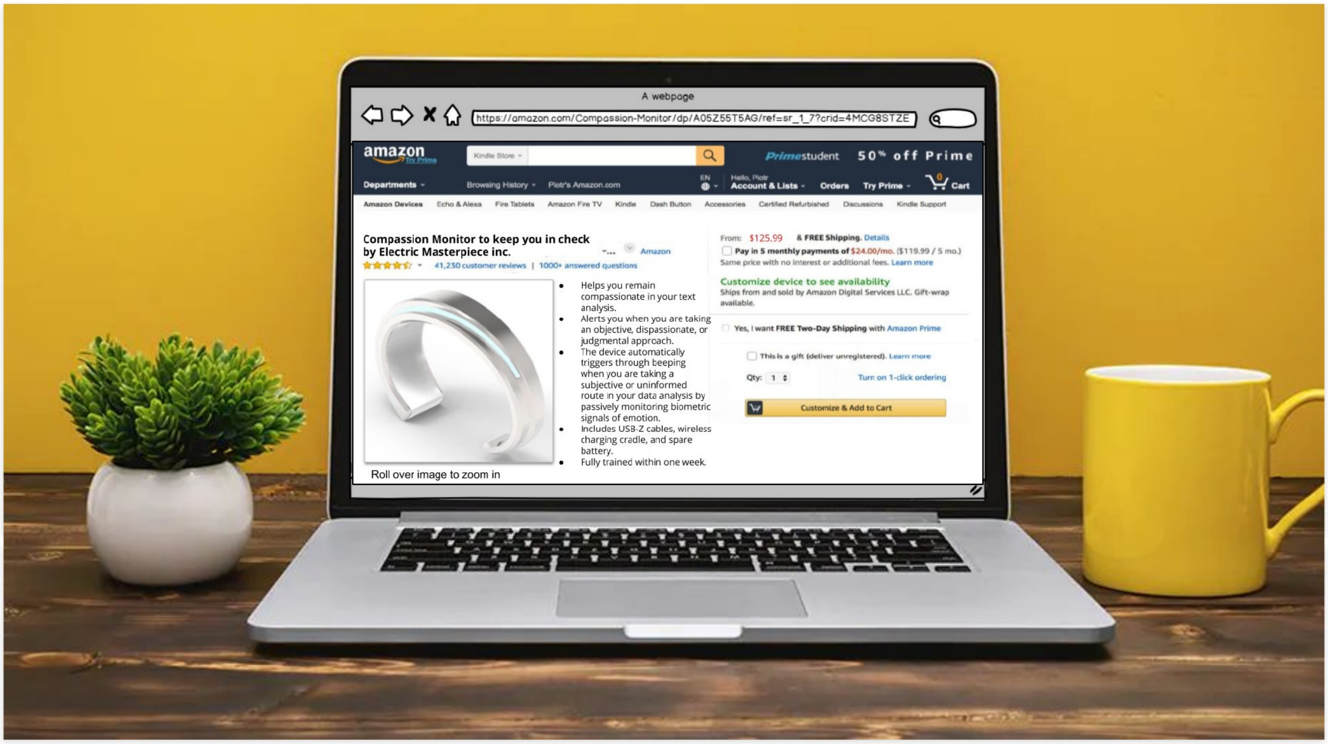
"I am able to empathize with the victims. It allows me to process legal documents faster because I don't have to go through thousands of pages of case trials to understand how the victim received justice."
Andan Macbe,
Legal Data Reporter

Figure 9: VRLegalPal imagined via user testimonials showing how a variety of users such as a legal data reporter have used and benefited from this design.



A wearable wireless sensor enabled device that pushes you to remain compassionate in your text data analysis task.

Figure 10: CompassionMonitor imagined as a landing page.



Compassion Monitor to keep you in check by Electric Masterpiece inc.

41,230 customer reviews | 1000+ answered questions

- Helps you remain compassionate in your text analysis.
- Alerts you when you are taking an objective, dispassionate, or judgmental approach.
- The device automatically triggers through beeping when you are taking a subjective or uninformed route in your data analysis by passively monitoring biometric signals of emotion.
- Includes USB-Z cables, wireless charging cradle, and spare battery.
- Fully trained within one week.

Roll over image to zoom in

From: $125.99 & FREE Shipping. Details

□ Pay in 5 monthly payments of $24.00/mo. ($119.99 / 5 mo.)
Same price with no interest or additional fees. Learn more

Customize device to see availability
Ships from and sold by Amazon Digital Services LLC. Gift-wrap available.

□ Yes, I want FREE Two-Day Shipping with Amazon Prime

□ This is a gift (deliver unregistered). Learn more

Qty: 1 Turn on 1-click ordering

Figure 11: CompassionMonitor imagined as a product on Amazon.