Interactive Model Cards: A Human-Centered Approach to Model Documentation

Anamaria Crisan*
Margaret Drouhard*
acrisan@tableau.com
mar.drouhard@tableau.com
Tableau Research & User Research
Seattle, Washington, USA

Jesse Vig
jvig@salesforce.com
Salesforce Research
Palo Alto, California, USA

Nazneen Rajani
nazneen@hf.co
Hugging Face
Palo Alto, USA

ABSTRACT

Deep learning models for natural language processing (NLP) are increasingly adopted and deployed by analysts without formal training in NLP or machine learning (ML). However, the documentation intended to convey the model's details and appropriate use is tailored primarily to individuals with ML or NLP expertise. To address this gap, we conduct a design inquiry into *interactive model cards*, which augment traditionally static model cards with affordances for exploring model documentation and interacting with the models themselves. Our investigation consists of an initial conceptual study with experts in ML, NLP, and AI Ethics, followed by a separate evaluative study with non-expert analysts who use ML models in their work. Using a semi-structured interview format coupled with a think-aloud protocol, we collected feedback from a total of 30 participants who engaged with different versions of standard and interactive model cards. Through a thematic analysis of the collected data, we identified several conceptual dimensions that summarize the strengths and limitations of standard and interactive model cards, including: stakeholders; design; guidance; understandability & interpretability; sensemaking & skepticism; and trust & safety. Our findings demonstrate the importance of carefully considered design and interactivity for orienting and supporting non-expert analysts using deep learning models, along with a need for consideration of broader sociotechnical contexts and organizational dynamics. We have also identified design elements, such as language, visual cues, and warnings, among others, that support interactivity and make non-interactive content accessible. We summarize our findings as design guidelines and discuss their implications for a human-centered approach towards AI/ML documentation.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → *Visualization*; **Human computer interaction (HCI)**; *Interaction design process and methods.*

KEYWORDS

model cards, human centered design, interactive data visualization

ACM Reference Format:
Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3533108

*Both authors contributed equally to this research.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533108

1 INTRODUCTION

Open source development has made it easier to share and deploy complex models, including large language models. This ease-of-use has lowered barriers to non-expert analysts [75] who do not have formal training in machine learning (ML), data science (DS), or linguistics. To accommodate a spectrum of ML expertise, Mitchell et al. [44] have proposed model cards as a means of providing consistent summaries of model details and their potential for misuse and harm. The format they propose is text-based and concise, making it both broadly accessible and applicable across model types. However, these model cards and similar forms of documentation rely on the developer to accurately and clearly report on the model and its performance. Often, this process is labor-intensive and many important details, such as unintended uses or disaggregated model performance, are omitted. Both experts and non-experts who want to interrogate the model further must do so by implementing their own analysis. Not only does this mode of interrogation leave many non-experts underserved, it also exacerbates the potential for harm once these models are deployed [5, 6, 11, 15].

Prior work has shown that non-expert analysts benefit from interacting with machine learning models and their data [1, 19, 59]. Recently, others have experimented with adding interactive elements to model cards. Both HuggingFace [73] and Google Cloud platforms introduced interactive modalities for interrogating the model's performance through customizable examples of the model's inputs. Robustness Gym reports [23] seek to overcome the development burden of model cards by allowing end-users to interactively create new slices of the data to interrogate a model's performance. Complementary to model cards are 'explainables', which are generally bespoke and interactive [30]. While these early explorations are promising, they do not explore how much or what kind of interactivity is beneficial.

In this work, we use the concept of a model card to scaffold a design inquiry into alternative and expanded forms of model documentation, focusing in particular on the needs of non-experts. We propose a novel concept for an *interactive model card* (IMC) that: (1)

lowers barriers to accessing key information about model behavior; (2) supports deeper interrogation of models; and (3) surfaces some attendant risks and limitations without additional work on the part of model developers. We built this idea out through an initial conceptual study with experts in ML, Ethics, and NLP to co-create a set of design guidelines for a functional IMC prototype. In this first study, we drew on experts' experiential knowledge of the wider implications of ML and NLP models, along with applications of these models within organizations, to *better understand implications of design choices in model documentation and inform our development of an IMC.* We conducted a subsequent study with non-expert analysts to *deepen our understanding of documentation needs and evaluate our IMC design.* Across both studies, we apply a human-centered lens to examine the affordances, opportunities, and limitations for standard model cards (SMCs) [44], Robustness Gym reports (RGRs) [23], and IMCs.

Through this research, we provide the following three contributions: 1) Design guidelines for interactive model cards and model documentation broadly; 2) A set of conceptual dimensions for evaluating model documentation with respect to AI/ML workers of various backgrounds; and 3) A functional prototype for an interactive model card that can be adapted for further inquiries and usage scenarios. We make artifacts of our research process available online[1].

2 BACKGROUND AND RELATED WORK

The ***potential for both broad and deep harm*** that AI/ML systems can pose to vulnerable people and ecosystems, especially when impacted people do not have the opportunity to interrogate and contest decisions driven by AI/ML(e.g., [4, 25, 26, 29, 72]). Large language models are especially predisposed to amplify harms, since they are being deployed widely enough to result in homogenization while their inner workings and limitations are poorly understood [5, 10]. Notably, even systems that seek to identify and mitigate racism, harassment and hate speech may compound harms to vulnerable people (e.g., [7, 14, 49]). In an effort to mitigate the risks for harm in AI/ML systems, researchers and developers have proposed strategies to make systems more ***explainable, transparent,*** **and** ***contestable*** (e.g., [3, 24, 34, 41, 45, 47, 74]). Our study design (Sections 3.2 and 4.2) has been informed by prior work related to measuring and improving model interpretability [45, 55], and proposals for contrastive explanations [45] are particularly aligned with our approach to interactivity and seeding the IMC with examples (as described in Section 4.1). Our approach to interactivity has also been informed by Amershi *et al.*'s guidelines for human-AI interaction [2], as well as by other guidelines and taxonomies for human-centered AI [3, 64] and proposals for enabling human contestation of AI-driven decisions [28, 40].

Prior work has also addressed dimensions of model behavior and documentation that impact ***trust*** and ***skepticism***. Many studies have found that cognitive biases (particularly anchoring bias and automation bias more broadly) can lead users to misunderstand model behavior and place unwarranted trust in AI systems [20, 34, 52, 55]. We introduce the concept of *productive skepticism* (Section 5), and we argue that stimulating productive skepticism–along with providing modalities for interaction and sensemaking–could be an effective strategy to mitigate the tendency to over-trust. Rather, opportunities to engage with productive skepticism can support stakeholders in calibrating appropriate trust in a model's behavior for a particular context and use case. Our approach aligns with calls for new strategies of calibrating user trust in AI [77] and promoting reflection [20], and our approach to guidance and warnings (discussed in Section 3.4) encompasses the requirements for extrinsic trust articulated by Jacovi *et al.* [33].

Various forms of ***model documentation*** have been proposed to make AI/ML systems more transparent and help users establish trust. *Model cards* (SMCs) [44] have been foundational in standardizing documentation for AI/ML model performance and characteristics, along with risks and unintended uses. Robustness Gym helps model developers report on model behavior through the generation of *robustness reports* (RGRs) that summarize model performance on various data slices [23]. We incorporate SMCs and RGRs as core elements of the IMC design. We also look to work on *datasheets for datasets* to include more detailed information of the model's data [21]. The IMC design follows the SMC model for reporting on datasets, and our functional prototype (Section 4.1) also includes warnings related to age of the training dataset. In evaluating the IMC, we also draw on dimensions articulated for *explainability fact sheets* [66].

Model cards and similar standardized artifacts can play a role in processes for ***governance*** and ***accountability*** of AI/ML systems. As many have studies have reinforced, mechanisms for governance and accountability of technical systems are prerequisites for informed consent [12, 22, 69] and equity [18, 37, 38]. Algorithmic audits are a key strategic approach toward governance and accountability. Whether conducted through internal processes or by third-party actors, audits seek to evaluate AI models for issues of robustness, bias, and fairness. Public audits have been shown to be effective at reducing bias in targeted companies (as compared to other companies) [56]. Auditing should be a structured, ongoing process that is designed from a system perspective [35, 64]. However, many challenges remain from conceptual, technical, economic, and organizational perspectives [46]. In contrast to posthoc audits and evaluations, *certificates* of robustness provide algorithmic guarantees on the performance of models under certain conditions [76]. Human-in-the-loop auditing processes leverage everyday users to provide further protection once a system is deployed [63], and we heard from participants in both of our studies that they believed IMCs would be valuable for record-keeping and organizational alignment throughout deployment.

Recent research has surfaced implications for ***practitioners' needs*** for trust and transparency. Some have identified strategies practitioners use to build understanding of AI/ML models, including data- and model-centric patterns of exploration [43, 61] and investigation of data where a model performs particularly well and poorly [39, 43, 61]. Our considerations around trust and productive skepticism have been informed by prior work identifying domain expertise and predispositions for trust as risk factors in bringing models into practice [17, 36, 43, 51, 75]. Furthermore, our support for exploration of sup-populations and sampling examples with low- and high-confidence scores address these needs in part. Hong *et*

[1] https://osf.io/9d83t

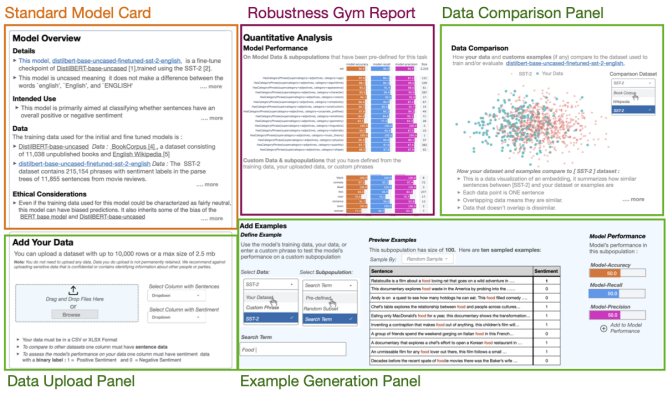

Figure 1: Concept of an Interactive Model Card. Larger images of the SMC, RGR, and IMC are available in the online materials

al. [32] have also surfaced needs for integrated interpretability support with AI/ML tooling, which aligns with our findings. In response to practitioners' needs to evaluate AI/ML models with respect to safety [71] and ethical decision-making [31], we have evaluated the IMC along these dimensions. As many researchers have articulated, safe and effective deployment of AI/ML systems in practice requires a robust understanding of different stakeholders' needs and goals for explainability [8, 32]; support for traceability and auditing [71]; and strategies to address concerns within existing organizational structures and dynamics [8, 31, 32, 57, 71]. We found that our approach to guidance, shareability, and traceability begins to address these needs, and we surfaced additional relevant implications for design.

3 STUDY 1 : CONCEPT STUDY WITH ML, AI, AND NLP EXPERTS

Ahead of engaging with non-expert analysts, we leveraged the expertise of researchers and practitioners in ML/AI development, AI Ethics, and Natural Language Processing (NLP), to co-create a set of design guidelines for IMCs.

3.1 Design Probes

Design Scope for Interaction. There are many ways that interactivity could be theoretically added to model cards. In this study, we have scoped the goals of interaction design to focus on disaggregated performance metrics, as well as understanding the model's training data and how it compares to the analyst's. We chose not to explore other sorts of interactivity, such as: comparing models (i.e., multiple sentiment analysis models), comparisons across multiple datasets, or training the model on the analyst's own data. We focus on disaggregated model performance and dataset understanding and comparison because we view these as key tasks for understanding a single model that are undersupported with an SMC. Moreover, understanding the model's performance and data conforms to what we consider the spirit of the SMC, which is to inform the reader about a single model.

Reference Model Card: Task, Model, and Data. We use a reference model card of DistilBERT [62] fine-tuned on the SST-2 [65] dataset for a sentiment analysis task; this reference mode is sourced from HuggingFace [73]. We chose to focus on a sentiment analysis task because it is widely known and easily understood, while still being representative of NLP model development regimens for other tasks. Moreover, the DistilBERT SST-2 fine-tuned model presented a unique challenge for generating design probes because it is technically three models: a pretrained general language model, its distillation, and the fine-tuned version optimized for a sentiment analysis task. We found that this challenged the SMC paradigm, since it was not clear how much information to include (if any) of the prior base models. To explore this problem, we decided to use the text from *both* the DistilBERT and SST-2 fine-tuned model cards[2].

Design Probes and Interactive Model Card Concept. We developed design probes [70] of an SMC, RGR, and IMC. For the SMC, we used the all the text from the reference model card, with some exceptions, laid out according to the specifications in Mitchell *et al.* [44]. Because the reference model card does have some interactivity (the ability to test your own sentence), we included that as well. An RGR unifies four common evaluation paradigms including performance on data subpopulations, which we explore in this concept study. We developed an RGR probe using the built-in subpopulations to which we added a custom set of potentially sensitive subpopulations (i.e., race, gender, etc.).

Our IMC concept (Figure 1) includes redeveloped elements of the SMC and RGR. We proposed adding interactive elements (shown with green border) that would allow users to upload their own data (`data upload panel`), add their own sentences (`example generation panel`), and/or define new subpopulations from the model's training and test datasets for exploration (`example generation panel`). We also include an interactive data comparison visualization (`data comparison panel`) that shows the sentence embeddings of the model's data together with the analyst's data. We implemented the design probes as paper prototypes using Google Slides. For the IMC we mocked up interactions through slide links, transitions, and animations. Text for these paper prototypes was taken verbatim, with some small exceptions, from the existing `DistilBERT` and the `SST-2 fine-tuned` model cards.

3.2 Study Procedures

3.2.1 Study Design. Our study comprised semi-structured interviews using a grounded approach [50] and incorporated conceptual design probes (Section 3.1). Study sessions were scheduled for 60 minutes, and participants were offered an honorarium of $150 upon completion. Full study protocols and materials are available in the online materials, while here we provide only a brief overview. During the study, participants were asked about their background and familiarity with model cards, Robustness Gym reports, or other types of model documentation. They were then shown examples of a standard model card, a Robustness Gym report, and our concept for an interactive model card (Section 3.1). They were asked to explore each of the forms in a think-aloud protocol, and the study moderator also prompted them with additional questions related to their interpretations and needs for understanding the model. As they considered each of these forms of model documentation,

[2] distilbert-base-uncased-finetuned-sst-2-english commit `03b4d19`

Table 1: Overview of roles and expertise of participants.

ID	Role	Expertise	SMC Familiarity	RGR Familiarity
E01	User Research Director	Ethics	No	No
E02	AI Ethics Lead	Ethics	Y es	Yes
E03	Ethics Data Scientist	Ethics	Yes	No
E04	Doctoral Student - CS	Auditability	Yes	No
E05	Doctoral Student - ML	Developer	No	No
E06	User Research Lead	Ethics	Yes	No
E07	Senior Researcher - NLP	NLP	Yes	Yes
E08	Postdoctoral Fellow - NLP/HCI	NLP	No	No
E09	Policy Research Fellow	Ethics	Yes	No
E010	Data Ethnographer	Ethics	No	No

we asked participants to reflect on the strengths and limitations through two lenses: their own needs as experts and those of a non-expert analyst using a sentiment analysis model in their work (see the online materials).

3.2.2 Recruitment. We set a target of 10 participants aiming to recruit at least one participant from each of the following categories: ML/AI development and design, Ethics, and NLP. We reached out to 23 participants, recruiting until we had met our target study size. Of the 13 participants who were not included in our study, five declined and the rest did not respond. We recruited participants based upon their publication or professional history using a combination of personal connections and cold emails. All participants had exposure to documentation for ML/AI models in some form, but only six were familiar with model cards and two with RGRs.

3.2.3 Data Collection and Analysis. Data collection included observer notes and video recordings for participants who consented to recording. We recorded approximately 11 hours and 15 minutes of video, including the interview and an on-camera debrief between the study moderator and note taker. Video transcripts were automatically generated, including speaker identification, and were verified for accuracy by the authors. We conducted iterative *thematic analysis* [13], surfacing initial codes through debriefs and review of transcripts. The first author then conducted focused coding for all transcripts, and both first authors synthesized themes and categories. All authors also surfaced implications that informed the design of our *IMC functional prototype* (Section 4.1).

3.3 Results

In Table 2 we summarize five interconnected conceptual dimensions of participant perspectives on our probes.

3.3.1 Stakeholders. Without prompting, participants described additional stakeholders or students and the scenarios in which they share model cards or other documentation (e.g., related to model performance and tuning). One participant succinctly summarized that SMCs remained relatively untested with non-expert analysts:

> *"At least for the next couple of years, people are going to be using model cards for the first time. And these are going to be concepts that may be new for many individuals. And so giving them some hand-holding, as to 'What does this mean, what am I supposed to do with this?' I think will be extremely helpful."* [E02]

Several participants (n=6) articulated that the language and layout of the SMC appeared better suited for software or machine learning engineers. One participant suggested that the SMC seemed *"very on the back end...like it feels like GitHub"* [E01], which they felt would not resonate with stakeholders who are using, but not implementing, models. Moreover, three participants reflected that some of these specialized terms of art (i.e., accuracy, precision, recall) can be difficult for even students of ML and NLP; one participant said their students needed a 'cheat sheet' of the terminology.

All participants expressed doubt that a non-expert analyst would understand the model card or know how to act on the information it contained. One participant suggested that an IMC could be more accessible to individuals across different roles, particularly if the design supported different information architectures or prioritization of information. In contrast, another participant [E07] was emphatic that standardization was a critical component of model cards and should be preserved. They argued that interactivity adds subjectivity to model cards that should be visually distinct.

3.3.2 Design Considerations. Participants critiqued the information design of all model cards; for the RGR and IMC, they also offered feedback on visual and interaction design. Most (n=6) participants identified overly technical language and jargon as a key failing of all of the model cards, and they proposed alternative language to capture the same meaning in more accessible ways. All participants said the amount of text in the standard model card made it difficult to quickly skim and understand the model. Most (n=8) participants expressed that the visual design and layout of the IMC helped their *"eyes [...] focus on information that [they] missed before"* [E04] compared to the SMC and RGR. However, a sense of focus/overwhelm varied among participants, and one participant articulated that the IMC contained *"too much information and is very cluttered"* [E04]. Four participants stated that a summary of the model's purpose (i.e., "this is a sentiment analysis model") was needed and that it would improve the readability of the model card. Relatedly, participants also pointed out areas in the RGR and IMC where they thought different information should be prioritized or emphasized (e.g., changing the color of the size column on the rightmost side of the RGR or emphasizing it in another way). Lastly, many participants suggested adding or emphasizing higher-level metrics about performance (e.g., making overall performance visually distinct from disaggregated performance).

3.3.3 Guidance. Participants indicated that various stakeholders would need help making sense of the information in the model card. Six participants emphasized the importance of providing specific guidance to help participants understand the information on the model card. Some forms of guidance suggested included providing definitions, or visual cues in the form of nudges and warnings. One participant articulated that explainers for visualizations would be helpful because *"it's pretty but..but what would I do [with it?]"* [E03]. One participant saw the direct potential for using interaction to provide guidance, suggesting that it *"would help me [...] dig deeper"* [E01] into the model details.

Other forms of guidance were complex and geared more toward *"help[ing] me understand why this matters [and] how does it help me make more informed decisions?"* [E02]. Supporting interpretability is complex but nearly all (n=7) participants indicated that examples could provide some useful guidance. However, having an effective default example to prompt end-users was also important [E04].

Table 2: Themes from our concept studies with experts in ML/AI, NLP, ethics, and auditability. Support quotes for each theme are also presented in the interview context from which they were elicited: SMC = Standard Model Card; IMC = Interactive Model Card; RGR = Robustness Gym Report; All = general comments not directed toward any specific model card type.

Theme	Subthemes	Example from interviews	
		Context	Quote
Stakeholders	Stakeholders, Language	SMC	*"Model cards look a lot more like a summary of an academic paper compared to getting a more general understanding of model."* [E07]
		SMC	*"Obviously, this, this standard model card is fantastic for an engineer who's coming in and trying to build that model [...] but [I am] not sure what their client needs."* [E09]
		All	*"There's a lot of jargon, some of it that can be inferred, but it's not something that is immediately apparent or made explicit"* [E10]
		RGR	*"[Precision, recall, accuracy], those are terms that are also going to be very confusing for someone with layman's background [...] I would also say that those terms are terms of art in many ways."* [E09]
Design Considerations	Information Design, Visual and Interaction Design, Implementation, Integrated Tooling	IMC	*[Hovering over points in the data comparison chart] I would want to see sentences because it breaks the abstract representation into some concrete things"* [E08]
		IMC	*"It is hard to discover the interaction, so make it clearer what you get with the interactivity and make it clearer that interaction is possible."*[E09]
		SMC	*"The developer has to make these model cards, but they are difficult to make, which is why these are not very informative."* [E08]
Guidance	Actionability, Defaults, Education, Explanations	All	*"People who aren't trained in ML don't necessarily know what they're looking out for or what questions to ask of both the data and the model [...] nudges could be useful."* [E04]
		All	*"Supporting analysis that says how to interpret this would be much more useful, but just even if it's formatted nicely with definitions, it's still not going to be super useful. Help me understand why this matters."*[E02]
		All	*"I wish that this [model card] would pull [me in more]. Tell me what to look at [...] maybe you could suggest the subpopulations that I might want to look at."* [E01]
		IMC	*"How much education about concepts do you need to bake in here? Do you have an intro in the beginning of this interactive model card?"*[E02]
Trust and Safety	Bias, Ethics, Misrepresentation	All	*"If I really am sort of coming in new to [model cards], I want to understand what my responsibility is [...] when I'm engaging with this model. I want to understand what agency I have."* [E01]
		SMC	*"Does this model card seem like it's showing a well-rounded representative model? Or is it one that has some big, ethical or transparent issues?"*[E03]
		All	*"Who gets to decide how to build those systems? And what is the end use case for evaluation? So in the healthcare space, specifically, a lot of times people get caught up with the area under the curve score F1 metrics for evaluating models about how they impact people's lives."*[E09]
Sensemaking and Skepticism	Contextualization, Data Analysis, Examples, Information Seeking, Interpretation	RGR	*"[I] might not pay as much attention to it [size of subpopulation]. And that really matters because you can see some of these examples–they have high scores, but then the [...] sample size is very small."*[E04]
		RGR, IMC	*"Just give [people] some example sentences and let them decide how they want to search or what they want to search for. And that way they can see model weaknesses and capabilities."*[E05]
		SMC	*"Some people read all the documentation straight; others, only when [...] they run into a challenge. And so there's a lot of information [in the standard model card] so, one question is [...] how many people are going to read through all this in advance when they're actually on the job and in practice?"*[E04]

Another participant suggested that leveraging examples with interactivity could help non-expert analysts *"build a better mental model of how the model works and what it lacks"* [E05].

3.3.4 Trust and Safety. Ethics and safety considerations were a top concern. The reference model card (Section 3.1) had a short ethical considerations text, but many participants took exception with the training dataset being described as "fairly neutral" (*"what does that mean even?"* [E08]). One commented that Silicon Valley tends to apply a standard set of *"moral values application across global and cultural contexts"* [E07], which diminishes the value of such ethical statements and assessments. Increasingly, such statements are seen as a *"cop-out because it just kind of says 'there is some bias somewhere' which is not entirely helpful."* [E06]. Two participants suggested stating explicitly whether bias assessments had been conducted, and one suggested raising a warning if such bias analyses are not run. Several (n=4) participants expressed that it could be beneficial to add sensitive populations to the disaggregated model performance, but they suggested caution since the definition of these groups was culturally and geographically contextual. Moreover, they emphasized the need to prominently display the total size of these subpopulations (and others) to support valid interpretation. For example, one participant initially thought the model performed well for a particular subpopulation, then saw the small sample size and revised their interpretation. Participants indicated that for these issues and overall, greater attention to information, visual, and interaction design would improve trust and safety.

3.3.5 Sensemaking and Skepticism. Some of the themes above converged as participants discussed the ways they believed non-expert analysts would explore and make sense of information in the model cards. Participants tied interaction positively to these themes, indicating that it would be beneficial to *"allow domain experts to be able to interrogate the models and formulate rules that are semantically meaningful to test against the model" [E08].* Importantly, the interactivity would enable people to *"elicit more questions"* [E04] that allowed them to form and test their mental models about the data and better *"see the things that model lacks"* [E05].

3.4 Summarizing Design Guidance for an Interactive Model Card

From our findings in this first study, we distilled a set of design guidelines (**DG**) to inform our functional IMC prototype. Based on what we learned from experts, these guidelines have broad implications for the documentation of AI/ML models and would be applicable beyond the task presented in the design probe.

DG1: Give careful deliberation to the design of information hierarchies, representations, and interactions. Prior work related to SMCs has focused primarily on the categories and depth of information that model documentation should include. However, to be sure that this content is effectively understood by a wide audience, the hierarchy and design of this content is equally important. Without this consideration, a model card may fail in its goals. Our conceptual study emphasized the importance of information design within both standard and interactive model cards. This includes the judicious use of language, visual cues, layout, data visualization, and interaction. Critically, our study cautions that the act of information design must be deliberate and, we argue, testable to be sure it is accurately interpreted.

DG2: Use interaction to help users develop conceptual understandings. The reference model card primarily used interaction as a preview of the model's outputs. However, experts saw the opportunity to use interaction as a way to help non-expert analysts develop their conceptual understanding of the NLP model. Interaction modalities must balance the subjectivity of contextual explorations with objective details about the model and its performance, so they should be designed and evaluated accordingly. The final consideration for interaction is that it must be discoverable and supported with defaults that help people orient to the affordances.

DG3: Scaffold important information with actionable guidance. Additional guidance will be necessary to support interpretability of the outputs. SMCs have been primarily tailored toward developers and machine learning experts, which can cause challenges for non-expert analysts who need to interpret the documentation for their own contexts. Clear guidance can support non-experts in understanding and knowing how to act on the information. Warnings, prompts, and summaries are all examples of guidance that participants suggested could help guide analysts in making sense of model cards. Visual cues can also support the overall architecture and flow of information in the model card.

DG4: Implement defaults that promote productive skepticism. Not all readers of a model card will want to interact with it. In addition to making interaction discoverable, it should also be enticing and promote a productive skepticism about what the model does and how it can be used. We argue that this productive skepticism would make ideas of intended and unintended uses more intuitive than simply being told what they are. Choosing appropriate default information is important to encourage this kind of engagement with a model card. We propose sensitive populations as a starting point and we used the Robustness Gym technology [23] to automatically implement these defaults.

4 STUDY 2 : EVALUATION WITH NON-EXPERT ANALYSTS

We used the design guidelines from the initial concept study to implement a functional prototype for an interactive model card (IMC) that we assessed with 20 non-expert analysts.

4.1 Interactive Model Card Functional Prototype

The functional prototype of the IMC reifies the design guidelines; in Figure 2 we show our IMC with themes from Table 2 overlain to emphasize the sources of design guidance. We visually separated information between a model overview component–reference model card information–and the 'contextual' component, where interactivity is used to probe the model's performance and underlying data. Within these components, we introduced elements of information and visual design to present the content (**DG1**) and create avenues of interactive engagement (**DG2**). Across both components, we use font faces, color, and highlighting to emphasize important content like the model's task (i.e., sentiment analysis) and its range of outputs. We also introduce two levels of information hierarchy. The first level summarizes vital information as a bulleted list; the second level provides technical details that are only visible when the analyst expands the content. In the overview component we also reduced the jargon from the SMC text.

We use interaction to support concept building, sensemaking, and skepticism around the model and its performance (**DG2**). The interactive data visualizations show the model's performance and explore the model's data. The performance visualization shows the overall train and test data set performance according to accuracy, precision, and recall (the reference model card only shows accuracy). Using Robustness Gym [23] we developed a set of terms that relate to subpopulations of US protected classes (e.g., race, gender, veteran status; **DG4**) that are also visualized We introduce three ways for analysts to further probe model performance: defining custom subpopulations within the model's data, adding their own sentences, and adding their own dataset. When an analyst chooses to add a sentence, they will see a summary of the model's sentiment prediction and have the option to contest the result; the 'sentiment label' for the new sentence is based on the participant's choice. We have also defined a set of default sentence templates of mixed sentiment [54] and with sensitive attributes [16] to help analysts further explore examples that are distinct from the model's data. Finally, the data visualizations are updated in real-time to reflect the model's performance in these newly defined subpopulations or data (sentence or dataset).

We scaffold the information in the model card with guidance that supports the interpretation of textual information and visualizations (**DG3**). Examples of guidance include simple statements explaining how to interpret visualizations or interact with them, definitions of terms, and instructions for adding new data. Lastly, we include a 'Guidance' section that includes more general information about sentiment analysis models and the interpretation of performance metrics. We incorporate warnings to alert the analyst to potentially unreliable or incomplete information in the model card. The performance visualization also alerts analysts when a subpopulation's size is lower than a specified threshold.

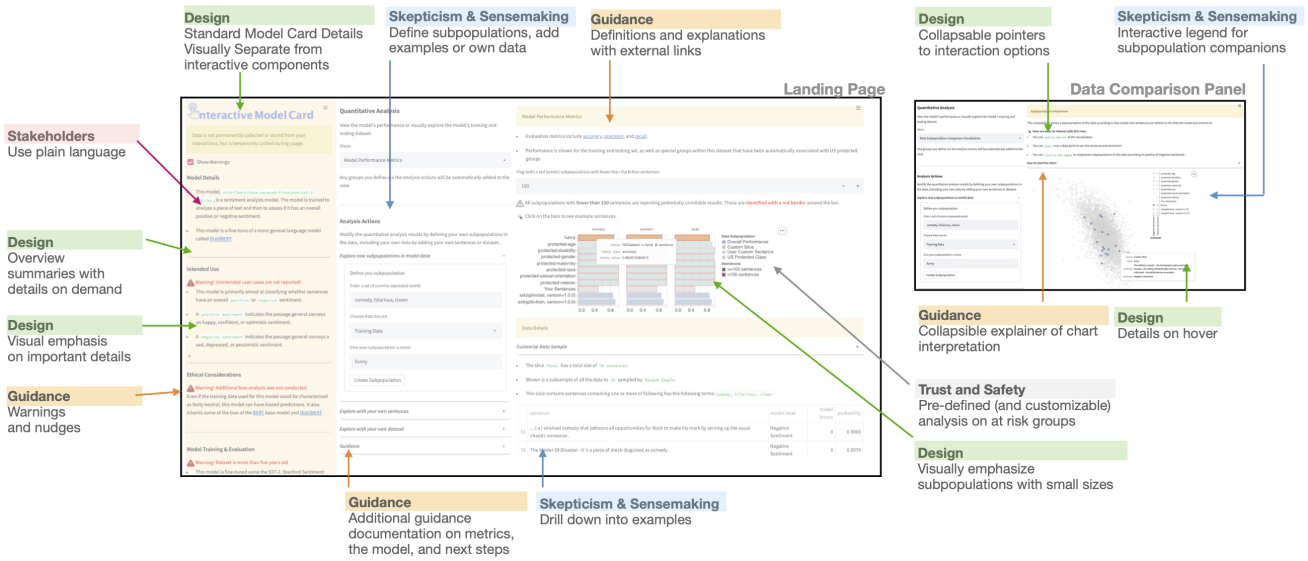

Figure 2: IMC Functional prototype that we presented to participants in the second study. Overlain are some examples of how feedback from the first study influenced the prototype design. The code for this prototype is available in the online materials.

The IMC is implemented in Python using `streamlit`[3], with visualizations and data interactions supported by `Altair` [68]. `Robustness Gym` [23] and `Gensim` [58] are primarily used to handle data inputs and outputs between the cards visual components, and are supported by other packages (`nltk` [9], `numpy` [27], `pandas` [42], `sklearn` [53]).

4.2 Study Procedures

4.2.1 Study Design. We conducted a 60 minute semi-structured interview with each of the participants, asking them to describe their use of ML or NLP models as well as the type of model documentation they consume or, if applicable, produce. Participants were then introduced to the format of a model card and directed to the DistilBERT SST-2 fine-tuned reference model card (Section 3.1). They were asked to explore this model card and think out loud about what information they would want to gather, whether it was available, and how they would wish to see this information. Next, participants were given a demonstration of the interactive model card and were asked to explore it as they had the reference model card. In the final 15 minutes of the session, participants were asked to consider the IMC with respect to its usability, functionality, safety, and ethics. These question prompts were developed based on the findings of the previous study and a set of explainability characteristics that we adapted from those described by Sokol and Flach [66]. We also asked participants to compare the IMC to the reference model card. At the conclusion of the study participants were compensated $125.

4.2.2 Recruitment. We set a target of 20 participants and recruited using the User Interviews platform. We screened eligible participants using a questionnaire and *a priori* inclusion and exclusion criteria. In total 176 participants responded. A total 117 participants were excluded, in some cases because they indicated never using ML or NLP models (n=89), while others (n=28) had graduate or undergraduate training in ML, statistics, or computer science. The remaining 59 candidates were separated into two groups according to the frequency with which they use code-based tools (i.e., Python, R, etc). Type I (n=35) analysts do not use programming tools at all or very infrequently (less than once a quarter), whereas Type II (n=24) analysts report daily or weekly use. We recruited an even number of participants from both categories of analysts (Table 3), anticipating each group might offer a different perspective to our study.Participants were compensated $125

4.2.3 Data Collection and Analysis. Data was collected and analyzed in the same manner as reported in Section 3.2. In this study we recorded approximately 23 hours and 30 minutes of video, which included both the interview and an additional on-camera debrief between the study moderator and notetaker. We used the themes surfaced in the first study to seed this analysis.

4.3 Results

The reference model card (Section 3.1), here acting as the SMC, and the IMC each had features that were preferred by both types of analysts. However, overall the IMC was favored over the SMC, with participants echoing the concerns that were previously articulated by the experts in Section 3. A summary of participant responses is shown in Table 4. Compared to the prior study, non-expert analysts provided greater depth towards the themes of *stakeholders*; *design considerations*; and *sensemaking & skepticism* in particular. This depth allowed us to expand on the subthemes from Table 2 and broaden our conceptions of how model cards might be used. In comparison to experts, non-expert analysts contrasted the *understandability & interpretability* of information between the SMC and IMC with more concrete details, leading to compelling insights around design implications. Additionally, participants' responses to our explicit prompts around usability, functionality, safety, and

[3]www.streamlit.io

Table 3: Overview of Participants for the Second Study.

Study Data		Background Details				Skills and Proficiency	
ID	Category	Region	Role	Industry	Org. Size	ML/DS Proficiency	Sentiment Analysis
A1-01	Analyst I	United States	Business and Marketing Analyst	Media	10000+	Basic	No
A1-02	Analyst I	Canada	Project Manager	Finance	5001-10000	Basic	No
A1-03	Analyst I	United Kingdom	Product Analyst	Insurance	201-1000	Limited	Yes
A1-04	Analyst I	United States	Finance Director	Finance	10000+	Limited	Yes
A1-05	Analyst I	United States	Product Analyst	Finance	10000+	Basic	No
A1-06	Analyst I	United States	Business Analyst	Information Technology	1001-5000	Advanced	No
A1-07	Analyst I	United States	Project Analyst	Environmental Services	1001-5000	Basic	No
A1-08	Analyst I	United Kingdom	Automation Developer	Information Technology	201-1000	Basic	No
A1-09	Analyst I	Canada	Business Data Analyst	Finance	5-200	Basic	Yes
A1-10	Analyst I	South Africa	Business Intelligence Analyst	Finance	1001-5000	Basic	No
A2-01	Analyst II	United States	Data Scientist	Government	10000+	Advanced	Yes
A2-02	Analyst II	United Kingdom	Analytics Manager	Marketing & Advertising	51 - 200	Advanced	Yes
A2-03	Analyst II	United States	Machine Learning Researcher	Material Engineering	1001-5000	Advanced	Yes
A2-04	Analyst II	United States	Machine Learning Developer	Information Technology	51-200	Advanced	No
A2-05	Analyst II	Canada	Business Data Scientist	Telecommunications	10000+	Advanced	Yes
A2-06	Analyst II	United Kingdom	Business Analyst	Finance	1001-5000	Advanced	Yes
A2-07	Analyst II	United Kingdom	Data Analyst	Information Technology	201-1000	Advanced	Yes
A2-08	Analyst II	United States	Data Engineer	Education	51-200	Advanced	Yes
A2-09	Analyst II	United States	VP Data & Analytics	Education	51-2000	Advanced	No
A2-10	Analyst II	Canada	Analytics Solutions Developer	Retail	1001-5000	Expert	No

ethics (summarized in Table 5) allowed us to expand upon dimensions surfaced in our first study. They also helped us contextualize participants' overall impressions of the IMC, summarized in Table 4. We did not observe large differences in patterns of responses between the two analyst groups. In Section 4.3.1, we outline participants' overall impressions of the IMC. After that we focus on three themes that highlight the importance of human-centered design in making model cards useful and usable in organizational settings.

4.3.1 Evaluating our Design Choices. Nearly all participants (n=18) found the IMC easier to read and understand (with some caveats discussed below). One participant said that the IMC showed information that *"allowed [them] to understand what this model actually is supposed to be used for"* and *"this is a lot better at level setting"* [A2-08]. Participants articulated that several components of the IMC helped them engage with the model and data more effectively by clarifying and prioritizing information, particularly: layout of information (n=13); language (n=6); ability to add examples (n=15); and visual and interactive elements (n=16). Three participants voiced the importance of being able to contest the model's results, and wanted to see this ability expanded beyond adding sentences (Section 4.1) to the model's training and testing data or uploading their own dataset. One participant stated that the ability to *"agree [or] disagree is kind of a check and balance"*[A1-06]. Nearly all (n=16) participants also made suggestions for improvements to the IMC design. These included changes to elements of the interface that were confusing (n=5), the layout (n=6), and types of guidance (n=1) provided. A few participants (n=4) expressed that text, visuals, and interaction options led to information overload. Participants suggested that integrating the model's code and files, present in the SMC, would be valuable for the IMC.

When asked, nearly all participants (n=15) indicated that the IMC was something they would use in their daily work, while some (n=2) indicated it was context-dependent. Participants reported that the IMC, with or without interactivity, was preferable to the SMC because the information was easier to understand. Notably, the addition of guiding elements, suggested in the previous study (Table 2), were well received and did help participants understand and interpret the model card information. Many of these non-interactive elements, which include visual cues, warnings, and prompts, among others, can be used to augment SMCs without having to add interactivity. However, participants indicated that interactivity added substantially to the SCM because it allowed participants to *"play with the data and see the results [in a way that] that feels super intuitive instead of explaining everything"* [A1-09].

4.3.2 Model Understandability - Getting the Basics Right. One of the most salient implications of the IMC design was that it helped participants accurately understand what the model was and what it did. In this study, we were surprised to discover that, when using the SMC, it was hard for participants to articulate *that the model performed sentiment analysis*. While experts had expressed that non-expert analysts would have some difficulties interpreting that information, this finding went beyond our expectations. Overall, almost half of the participants were not able to provide the correct interpretation of what the model did using the SMC alone. Among these participants, four thought it was a general text classification model, but could not anticipate its outputs (positive or negative sentiment classification), two thought that DistilBERT and its fine-tuned version were both sentiment analysis models, one participant did not answer (*"I would have a tough time understanding what it does."*[A2-08]), one thought that SST-2 dataset was the model but could not articulate what the model did. Even three participants who did correctly describe the model were not confident in their interpretation : *"I would guess that it's sentiment analysis based on text, but I really don't know"*[A1-03].

Table 4: Themes from our evaluative studies with non-expert analyst. The reference model card represents the SMC in this study.

Theme	Subthemes	Example from interviews	
		Context	Quote
Stakeholders	Community, Language, Sharing, Teams	SMC, IMC	*"Overall I am really impressed with the language clarity that is there right now, compared to [the reference model card]; the biggest difference is its clarity."* [A2-09]
		IMC	*"We actually have had a lot of communication issues in the team between our data science guru and the rest of the business and I think this actually might be really helpful"*[A2-02]
		IMC	*"If you're trying to bring in a machine learning tool or a machine learning process, you've got to go through so many hoops to get that sign off, this is a sort of thing that would just be a godsend to have"* [A1-08]
		IMC	*"This [IMC] is something that I could send in a pinch to any stakeholder. So if I was in a meeting with a whole bunch of stakeholders [...] and somebody's like, 'Hey, what is the model? What does it do?' You know, I could answer most of these questions. But in a pinch, I'd say like, here's a link, go interact with it."* [A2-04]
Design	Assessment & Validation, Comparison, Defaults, Information Priority, Interaction, UX Patterns, Visualization	SMC	*"The first thing I am looking for is an introduction to the model and what it does and what it is used for. I don't know if I see it here"* [A2-10]
		IMC, SMC	*"So, interactivity is the main one. This is more of an all or one, whereas the [SMC] makes me use another tool. It [SMC] is more like a shopping mall, where I need build it [the analysis] on my own at home."* [A1-03]
		IMC, SMC	*"You want it all accessible, the fact that it's one page is great. Like with the other one I had to click around a lot to get the information, but this model card doesn't have that problem, which is great."* [A2-02]
		IMC	*"It gives me a good outline of what I am looking at. The few examples, short, sweet, and to the point are helpful."*[A1-07]
		IMC	*"Maybe it's my time working in finance, but bullet points work for me"* [A2-02]
Sensemaking & Skepticism	Awareness, Confidence, Contestability, Examples, Exploration, Prompts, Warnings, Risk Group Analysis	SMC, IMC	*"It [SMC] seemed more oriented towards somebody who wants to, like develop and train it [the model] and deploy it [...] this [IMC] does give me the same information, like it gives me the same context [...] but it's also giving me the tools to kind of look at the model on a higher level."* [A2-04]
		IMC	*"I absolutely like this quantitative analysis [...] which lets me play around, and I can see whether the model is working or not. So I'm not just taking the the word of the developer of the library"* [A2-09]
		SMC, IMC	*"if I ever see something that just reports one metric [...] I'm always like a little bit skeptical that maybe they just picked that up because that's what looks good"*[A2-01]
		IMC	*"I do like the fact that you can [...] agree or disagree [with the model's prediction], because sometimes it's nice to have that ability* [A1-06]
Understandability & Interpretability	Accessibility, Clarity, Data Understandability, Guidance, Model & Result Interpretation, Information Overload, Information Sufficiency, Trust	IMC	*"This gives me more assurance that this is the right model or not"* [A2-09]
		SMC	*"I can tell it, it's used for some text classification[...] But, in terms of what this thing does, it's not obviously intuitive enough, unless you've caught some context around it. So I filled in a lot of the blanks."* [A2-08]
		SMC	*"What is the sentiment treebank [SST-2 dataset]? Like, is it a list of sentences? Where did they come from? Who generated them? Because that was also what I was confused"* [A2-01]
		SMC	*I feel that it's too much information on this model. If, maybe, could I have a better summary?* [A1-10]
		SMC	' *"The most important bit is what the purpose of the model is, and it's good at the top, otherwise I didn't know what what I was looking at, I had to guess how it actually works and what is trying to tell me"* [A2-05]

One source of confusion was the need to consume information across two model cards (DistilBERT base model and SST-2 fine-tuned). Guided by our earlier study, we tackled this challenge directly and used the design elements of the IMC to draw attention to what the fine-tuned model did and how it differed from the base model. The response of participants validated this approach, with one stating that *"the side panel was exactly what I was looking for and it highlighted some of the words, then I know that this is something that I might be interested in looking at"* [A2-06].

4.3.3 The Role of Model Cards in Sociotechnical Systems. Nearly all (n=18) participants described, without being prompted, how the IMC would help them share information with internal and external stakeholders. The actionability of a model card was frequently (n=13) tied by participants to its ability to be shareable and understandable to others with whom they work. Moreover, participants discussed interwoven social and technical processes and described where either an IMC or SMC would fit within them. Participants described how the model details (overview component in the IMC, Section 4.1) could help them with an initial assessment of the model's validity for their use case. A quick answer is essential, and one participant stated, *"I would lose patience after 30-40 seconds if I have to put a lot of effort into what I'm looking for"*[A2-06]. If the participants were convinced the model was suitable for their use case, they would move on to the interactive components to explore it further. One benefit of this interactivity was that it allowed participants to conduct lightweight experimentation, which was cumbersome using the SMC because *"you [had to] deploy the model in Python yourself, put in your own sentences"* whereas with the IMC *"just being able to do it like this...ease of use is off the charts"* [A2-02]. Finally, the model card would be used to start a dialogue with other stakeholders. These stakeholders included executive or technical team members, including data scientists.

Beyond integrating model cards with their social organization, analysts (n=7; the majority belonging to group II) saw the benefits of incorporating model cards into their existing technical infrastructure: *"A huge win would be getting this to integrate seamlessly with open-source frameworks"* [A2-01]. Here, the reference model card contained information and actions to obtain the model code and launch it — something that was absent in the IMC. Moreover, three participants expressed that IMCs, if they could be easy to produce, would lower their existing model documentation burden.

Table 5: Examples of participant responses on the usability, functionality, safety, and ethics of the IMC.

Prompt	Definition	Example from interviews
Usability	The efficacy of the design choices to help assess and digest the information in the model card	*"This feels like plain English [...] like a very simple way of saying, 'hey, this is what we have. this is what it does.' "* [A1-01]
		"I think this [IMC] tells a better story, [...] and this tells me, what's the model? What's it supposed to be used like?" [A2-08]
		"It's not only the numbers, but it's also the insights and visually being able to see[...] what those numbers are saying" [A2-04]
		"The visualization, as far as the reporting goes, it has got a good description of what all that means. A lot of the help text and description text is there." [A1-06]
Functionality	The ability to use the IMC to assess the relevance and applicability of the model to their routine work	*"This would be perfect for internal stakeholders, like technical stakeholders, direct managers, CTO, data science team, you know, people like that. I would want a simplified version of this for external stakeholders. But I also have all the data here and it looks like I can extract it if I want to."* [A2-04]
		"If I had a tool like this, I would start to explore, and pick out trends or phrases, in their data, so that I can start to see that there are opportunities from a features development perspective."[A1-03]
		"I've got a data set that I want to run sentiment analysis on, I'm going to use this model, drop it in through my CSV, boom, now I want to see the visualizations great. And they're going to look really good when I'm presenting to my boss."[A1-03]
		"It [the model card] would be more about ruling out that it's not the right model, you know, and then I might have like six or seven model cards open, as I'm going through all the models, I can find, like, here's my list of things we should try."[A2-07]
Safety	The ability to assess safety risks for a model, including risks for security, privacy, robustness, and related dimensions	*"I think it [the IMC] gives you the basis to have that conversation."*[A2-02]
		"It is the person's responsibility to make sure they know what they are doing and not someone who is providing the model to them." [A2-06]
		"The safety risk is more in the problem itself and not the model you are applying." [A2-07]
		"With the protected classes I think it's a good feature to have, I think it's one of the more important ones. Protected classes make me think that things are headed in the right direction. Also I like the warnings." [A2-08]
Ethics	The ability to assess ethical consequences encompassing things like potential harm to vulnerable people or ecosystems, unintended bias, or similar issues	*"I think it's useful as well to say, 'You know, what, you're gonna have to go and do your own research' [...] you can't just say 'Oh, just go and read this.'[Use the] model card to assist you, but you're going to have to go and do a bit of learning yourself."* [A1-08]
		"I don't think it helps me with ethical consequences, but it helps to give me some considerations." [A1-03]
		"I think also having these categories [protected classes] made me curious and think about it." [A1-09]
		"Bias is also like a term of art in machine learning, right? So it's like, I read it as bias, like, Oh, it's just talking about bias from the standpoint of overfitting. It has nothing to do with like, discrimination and ruining people's lives."[A2-07]

4.3.4 Ethics and Safety are Challenging Topics. Participants had difficulty contextualizing the safety and ethical dimensions of either the SMC or IMC. Although the IMC included templates to examine ethically challenging examples (Section 4.1), only four participants expressed that these features could stimulate discussion. The unintended uses of the model were similarly inaccessible, with one participant stating that *"unintended uses are important [but] how would I know what [is] unintended?"* [A1-01]. One participant expressed doubt that any model documentation could really give an adequate picture of ethical and safety risks, emphasizing that ' *'this [model card] would not be my only data point or else, you know, that's a very quick way to lose your job"* [A1-04]. Although the goal of model cards is not to be a substitute for further analysis of ethical and safety risks, this participant's comment echoes earlier observations from experts (Section 3.3.4) that information about ethical risks is often dismissed. Others felt that topics of safety and ethics were too abstract, with one participant stating that they *"wouldn't worry too much about that to be honest, I guess it depends on the type of analysis you're doing."* [A2-05]. While the IMC also provided template sentences to promote ethical thinking, participants largely favored their own examples. Unexpectedly, two participants pushed back against the inclusion of more actionable ethical and safety content, articulating concerns that too much guidance could be detrimental and provide a false sense of security.

4.4 Summary

We assessed our functional prototype for an interactive model card with twenty non-expert analysts drawn from across industry. Their feedback validates many of our design choices, but also offers avenues for further improvements. These non-expert analysts also illuminated a more detailed view of how model cards, more generally, might be used within the sociotechnical systems in their organizational contexts. Our findings point to the importance of information and visual design for transparency and interpretability of model cards.

5 DISCUSSION

Technology that enables low or no code data analysis is lowering barriers to developing and deploying deep learning models. As E02 astutely observed (Section 3.3.1), the net effect is to broaden the field of stakeholders who will need access to model cards in the future. Increasingly, the data workers taking on AI/ML development do not have formal education in the theoretical and technical underpinnings of these systems. We posited that, for this group, augmenting model cards with interactivity would make AI/ML systems more interpretable.

We recruited and explored the perspectives of people who we believe are representative of the future data workforce. Our findings show that interaction in model cards helped participants better understand model behavior and implications for their work. Critically, the benefits of interaction were realized primarily through the careful architecture of information that included the choice of data visualizations, layout, language, visual cues, warnings and prompts. We contend that without these careful design choices, model cards are inaccessible to many individuals who use AI/ML models in their work. Our studies have also surfaced the value proposition data workers identify for model cards to support organizational

decision-making and traceability. From our collective findings we draw four key insights:

1) Interactive model cards are a bridge to further analysis, not a substitute. How interactive should a model card be? At what point does it cross over into an analysis interface? In designing the IMC, we wrestled with these questions, and study participants were divided on the appropriate balance. We argue that with thoughtfully designed interaction, model cards can be a first step for data workers to explore how trustworthy a model is and whether it might meet their needs, serving as a guide to deeper interrogation when desired. For individual stakeholders, the most effective form for interactivity in model cards will be situated and contextual. However, future research could deepen our understanding of how particular design decisions impact the understandability and actionability of model cards.

2) Interactivity and guidance bolster productive skepticism. Interaction supports individuals in contextualizing and interrogating model behavior (e.g., by adding their own data or contesting the model's results). The opportunity to *"be really curious to just kind of play around"* [A2-05] can help mitigate anchoring bias [52]. This sort of interactivity may encourage *productive skepticism,* an orientation that is neither overly dismissive nor trusting. Importantly, we found that scaffolding these interactions to support sensemaking requires clear and actionable guidance. Our research has only scratched the surface of the dimensions and forms of guidance that help people understand and act on model behavior. There remain many avenues to explore with IMCs as a foundation for this inquiry.

3) Model cards should support the data work community, not just an individual. Decision-making with data and deep learning models is a collaborative and distributed process involving information sharing over time and across organizational roles. For model cards to be adopted and impact organizational processes, they must support knowledge sharing and negotiation across stakeholders with diverse backgrounds and perspectives. Furthermore, the creation of interactive model cards need not be so burdensome. Our findings suggest that collaborative development, along with integrated programmatic tooling, could lower the burdens for model card generation, reuse, and refinement.

4) Unintended uses, ethics, and safety are too ambiguous to be actionable. Model cards are intended to surface, among other things, ethical and safety implications – a topic that is especially pressing for deep learning models. However, our research supports prior findings that these concepts are challenging to integrate into decision-making [48, 60], and we identified some cynicism toward ethics statements, as well as assertions that ethical questions require contextual, situated examination. Our findings indicate, unsurprisingly, that model documentation can spark ethical thinking, but will never substitute for it. We encourage further inquiry into other strategies toward fairness (e.g., metrics [16, 67]), and we urge caution to ensure that these approaches are understandable and actionable for non-experts. Importantly, our analysis reinforces that considerations of ethics are personal and contextual, so they are unlikely to be addressed with a single approach.

Limitations and Future Work. Design choices and the choice of sentiment analysis tasks (Section 3.1) likely influenced our findings. We anticipate that specific IMC design elements (i.e., the choice of visual cues, data visualizations, etc.) will be refined and adapted in future work. However, we believe that the broader design guidelines (Section 3.4) and conceptual dimensions that we surfaced (Tables 2, 4, and 5) are robust and will generalize to other tasks and contexts. We also recognize that creating and learning how to use IMCs is a heavier burden than for text-based model cards. We have taken steps in this work to build toward adaptable and extensible model card generators. Future work will explore further strategies for reducing and re-distributing the work of model documentation.

6 CONCLUSION

Together with ML/AI experts and non-experts, we co-developed a concept and functional prototype for an interactive model card for a large language model. Within the broader range of stakeholders using and deploying NLP and other ML/AI models, our work presents a timely and important examination of model cards from a human-centered design perspective. We foresee future opportunities for other researchers, practitioners, and developers to build from our findings to ensure that ML/AI models are interrogable, contestable, and documented responsibly.

ACKNOWLEDGMENTS

We wish to thank Britta Fiore-Gartland, Anna Bethke, Kathy Baxter, Wenhao Liu, Emily Witt, and Vidya Setlur for their thoughtful feedback on this work. We also gratefully acknowledge study participants for their time and insights.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Mag.* 35 (2014), 105–120.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. *Guidelines for Human-AI Interaction.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *Calif. L. Rev.* 104 (2016), 671.

[5] 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.

[6] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366 (2019), 421 – 422.

[7] Sebastian Benthall and Bruce D Haynes. 2019. Racial Categories in Machine Learning. In *Proceedings of the conference on fairness, accountability, and transparency.* 289–298.

[8] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20).* Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[9] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[10] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2021).

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[12] Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, et al. 2020. The CARE Principles for Indigenous Data Governance. (2020).

[13] Kathy Charmaz. 2014. *Constructing Grounded Theory*. sage.

[14] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).

[15] Anamaria Crisan and Brittany Fiore-Gartland. 2021. *Fits and Starts: Enterprise Use of AutoML and the Role of Humans in the Loop*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445775

[16] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. arXiv:2106.14574 [cs.CL]

[17] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology. General* 144 1 (2015), 114–26.

[18] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. 2016. Social Justice-oriented Interaction Design: Outlining Key Design Strategies and Commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. 656–671.

[19] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (jun 2018), 37 pages. https://doi.org/10.1145/3185517

[20] Upol Ehsan and Mark O Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[22] Alexandra Giannopoulou. 2020. Algorithmic Systems: the Consent is in the Detail? *Internet Policy Review* 9, 1 (2020).

[23] Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness Gym: Unifying the NLP Evaluation Landscape. *arXiv preprint arXiv:2101.04840* (2021).

[24] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[25] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13.

[26] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.

[27] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585 (2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[28] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 95–99.

[29] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.

[30] Fred Hohman, Matthew Conlen, Jeffrey Heer, and Duen Horng (Polo) Chau. 2020. Communicating with Interactive Articles. *Distill* (2020). https://doi.org/10.23915/distill.00028 https://distill.pub/2020/communicating-with-interactive-articles.

[31] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[32] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.

[33] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[34] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[35] Sara Kingsley, Clara Wang, Alex Mikhalenko, Proteeti Sinha, and Chinmay Kulkarni. 2020. Auditing Digital Platforms for Discrimination in Economic Opportunity Advertising. *arXiv preprint arXiv:2008.09656* (2020).

[36] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 141–148.

[37] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 772–781. https://doi.org/10.1145/3442188.3445938

[38] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[39] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 90–98. https://doi.org/10.1145/3351095.3372824

[40] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 106 (apr 2021), 25 pages. https://doi.org/10.1145/3449180

[41] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc Interpretability for Neural NLP: A Survey. *arXiv preprint arXiv:2108.04840* (2021).

[42] Wes McKinney et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. Austin, TX, 51–56.

[43] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[44] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[45] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 279–288. https://doi.org/10.1145/3287560.3287574

[46] Jakob Mokander and Luciano Floridi. 2021. Ethics-based Auditing to Develop Trustworthy AI. *arXiv preprint arXiv:2105.00002* (2021).

[47] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable Machine Learning–A Brief History, State-of-the-Art and Challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 417–431.

[48] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2019. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. arXiv:1905.06876 [cs.CY]

[49] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model. *PloS one* 15, 8 (2020), e0237861.

[50] Michael Muller. 2014. Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method. In *Ways of Knowing in HCI*. Springer, 25–48.

[51] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[52] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proc. IUI'21*. 340–350. https://doi.org/10.1145/3397481.3450639

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[54] Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. DynaSent: A Dynamic Benchmark for Sentiment Analysis. (2020).

arXiv:2012.15349 [cs.CL]

[55] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[56] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[57] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 7 (apr 2021), 23 pages. https://doi.org/10.1145/3449081

[58] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[59] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. 2017. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* 268 (2017), 164–175. https://doi.org/10.1016/j.neucom.2017.01.105 Advances in artificial neural networks, machine learning and computational intelligence.

[60] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. *"Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445518

[61] Téo Sanchez, Baptiste Caramiaux, Jules Françoise, Frédéric Bevilacqua, and Wendy E. Mackay. 2021. How Do People Train a Machine? Strategies and (Mis)Understandings. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 162 (apr 2021), 26 pages. https://doi.org/10.1145/3449236

[62] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

[63] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 433 (oct 2021), 29 pages. https://doi.org/10.1145/3479577

[64] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.

[65] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proc. EMNLP'13*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://aclanthology.org/D13-1170

[66] Kacper Sokol and Peter Flach. 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 56–67. https://doi.org/10.1145/3351095.3372870

[67] Rachel Thomas and David Uminsky. 2020. The Problem with Metrics is a Fundamental Problem for AI. *arXiv preprint arXiv:2002.08512* (2020).

[68] Jacob VanderPlas, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. Altair: Interactive statistical visualizations for python. *Journal of open source software* 3, 32 (2018), 1057.

[69] Salome Viljoen. 2021. A Relational Theory of Data Governance. *The Yale Law Journal* 131, 573 (2021).

[70] Jayne Wallace, John McCarthy, Peter C. Wright, and Patrick Olivier. 2013. Making Design Probes Work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(Proc. CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3441–3450. https://doi.org/10.1145/2470654.2466473

[71] Jennifer Wang and Angela Moulden. 2021. AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[72] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018.* AI Now Institute at New York University New York.

[73] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL]

[74] Jilei Yang, Diana Negoescu, and Parvez Ahammad. 2021. Intellige: A User-Facing Model Explainer for Narrative Explanations. *arXiv preprint arXiv:2105.12941* (2021).

[75] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 573–584.

[76] Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. 2021. Certified Robustness to Programmable Transformations in LSTMs. *arXiv preprint arXiv:2102.07818* (2021).

[77] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852