Evaluating a Methodology for Increasing AI Transparency: A Case Study

DAVID PIORKOWSKI, IBM Research AI, USA
JOHN RICHARDS, IBM Research AI, USA
MICHAEL HIND, IBM Research AI, USA

In reaction to growing concerns about the potential harms of artificial intelligence (AI), societies have begun to demand more transparency about how AI models and systems are created and used. To address these concerns, several efforts have proposed documentation templates containing questions to be answered by model developers. These templates provide a useful starting point, but no single template can cover the needs of diverse documentation consumers. It is possible in principle, however, to create a repeatable methodology to generate truly useful documentation. Richards et al. [25] proposed such a methodology for identifying specific documentation needs and creating templates to address those needs. Although this is a promising proposal, it has not been evaluated.

This paper presents the first evaluation of this user-centered methodology in practice, reporting on the experiences of a team in the domain of AI for healthcare that adopted it to increase transparency for several AI models. The methodology was found to be usable by developers not trained in user-centered techniques, guiding them to creating a documentation template that addressed the specific needs of their consumers while still being reusable across different models and use cases. Analysis of the benefits and costs of this methodology are reviewed and suggestions for further improvement in both the methodology and supporting tools are summarized.

CCS Concepts: • Software and its engineering → Documentation; • Computing methodologies → Artificial intelligence; • Human-centered computing → Field studies.

Additional Key Words and Phrases: AI Transparency

1 INTRODUCTION

AI-based systems are increasingly being used to make or inform decisions that can greatly impact an individual’s life. Although these systems can be beneficial, numerous groups have demonstrated significant risks of these systems [5, 6, 12, 16, 26]. Due to these concerns, there has been strong demand for increased AI transparency from researchers, civil society advocates, and governments. One way to achieve this transparency is to have a standardized document, similar in spirit to a nutritional label, that would describe the relevant aspects of the AI system. Several research [1, 3, 11, 17, 18, 24], industry [10], and government [7, 19, 28] efforts have proposed documentation templates containing questions to be answered by model or dataset developers. Although these templates provide a useful starting point, no single template can cover the needs of diverse documentation consumers. These include people not involved with the development of the AI model: e.g., end users, affected subjects, and regulators, as well as those who have developed and deployed the model: e.g., data scientists, model validators, and model operations engineers.

While a universal “AI nutritional label” may be appropriate for more general transparency goals, it will leave gaps due to the specifics of the use case, model implementation, problem domain, and regulatory context. To address these gaps, we can strive for a consistent, repeatable, flexible process for creating consumer-appropriate transparency documentation. Richards et al. [25] proposed such an approach, following the principles of user-centered design, to identify the needs of both producers and consumers of AI documentation in the form of a methodology to create AI FactSheets [1]. Although promising, it has yet to be evaluated with real AI systems.
The FactSheets methodology is a documentation process centered on three roles: content producers, content consumers, and a FactSheet team (FS team). The FS team follows the methodology and shepherds the creation of a FactSheet. During this process, they collaborate with both producers, who have the information to be captured, and consumers, who have specific documentation needs. The FS team determines consumers’ needs and records them as fields to be completed in a FactSheet template. Then, the FS team iterates with producers and consumers to fill out and refine the template, producing a FactSheet.

The goal of this paper is to evaluate this methodology via a case study of an independent team that built FactSheets for multiple real AI models and systems over three months. We interviewed 16 participants to address the following research questions:

RQ1 Is the methodology usable by FS team members not trained in human-centered practices?
RQ2 How well did the resulting FactSheets address the needs of different consumers?
RQ3 What did consumers and producers of FactSheets see as the benefits and costs?

2 BACKGROUND
AI transparency requires the availability of model and system documentation that is understandable and trustworthy. Documenting any software artifact well requires effort. AI models, and the systems in which they are embedded, pose additional documentation challenges [13]. First, AI development is highly collaborative. People with diverse skills, specialized vocabularies, and unique tools all contribute to the final deployed model or system. Second, AI development is highly iterative. Model development often begins with a series of lightweight experiments. Multiple threads are followed with frequent false starts and backtracking. Along the way, important decisions about training data and algorithm refinement are rarely captured. Finally, tooling to support traditional software development is not easily adapted to AI development and documentation. Proposals for new mechanisms for collecting important facts about AI development throughout the lifecycle [13], may lead to improved tooling in the future. Currently, only the most disciplined teams can reliably capture information needed for complete and accurate documentation.

2.1 FactSheets and FactSheets Methodology
To tackle the complexity of creating AI documentation that affords effective transparency, Richards et al. [25] have proposed a human-centered methodology for creating the form of documentation called AI FactSheets. FactSheets [1, 13, 14] are collections of important facts about the development, testing, and deployment of AI models and systems. These include facts about model purpose, training data selection and cleaning, algorithm selection and tuning, testing for accuracy, bias, privacy risks, adversarial attacks, etc. Different consumers of FactSheets (data scientists, business owners, system integrators, deployment engineers, risk officers, regulators, end users, affected subjects, etc.) will have different skills and different information needs. By engaging with these consumers, it is possible to discover both common and unique needs and, thereby, create FactSheets that meet those needs. Those needs may include how to render the documentation, providing for example, a condensed table view for a quick overview or slides to aid presentation of the model [14]. By engaging with fact producers, it is also possible to discover what facts they can create for consumption by others and how to help them create these facts efficiently. Earlier work [25] has posited that executing the steps of this methodology will lead to useful FactSheets, contributing to model transparency.

The methodology consists of seven steps. The following list presents a somewhat idealized view insofar as there may be iterations within and between steps. In other cases, some of the steps may be collapsed due to informants playing
both fact consumer and fact producer roles within the lifecycle [13]. Further details on these steps, including example questions to explore with informants, can be found in [25].

(1) **Know Your FactSheet Consumers.** FactSheets (and by extension, any form of AI documentation) are produced so that they can be consumed. Understanding the information needs of FactSheet consumers is the first and most important task. This initial exploration need not be formal. Working with even one representative informant from each major process in the AI lifecycle will provide useful insights into the overall set of consumer needs.

(2) **Know Your FactSheet Producers.** Some facts can be automatically generated by tooling. Some facts can only be produced by a knowledgeable human. Both kinds of facts will be considered during this step. Again, working with even one representative producer of relevant facts from each major process in the AI lifecycle will provide the information needed to proceed to the next step.

(3) **Create a FactSheet Template.** What is learned in the first two steps leads directly to the most important part of creating FactSheets, namely the creation of a FactSheet template. A FactSheet template will contain what can be thought of as questions or, alternatively, the fields in a form. Each individual FactSheet will contain the answers to these questions. For example a template may start with the question “What is this model for?”. It may then expand on that question by asking where the model is well-suited and where the model is ill-suited.

(4) **Fill In FactSheet Template.** This step is where the creator or creators of the FactSheet template attempt to fill it in for the first time. As this is done, it is important to informally assess the quality of the template itself by reflecting on what was learned about both consumers and producers, their skills, and information needs in the first two steps. While this assessment is not a substitute for further work with them (to follow), it may quickly highlight where improvements are needed.

(5) **Have Actual Producers Create a FactSheet.** In this step, actual fact producers fill in the template for their part of the lifecycle. For example, if there is a question in the template about model purpose, find someone who would actually be providing that information and have them answer the question. Ask a data scientist to answer the questions related to the development and testing of the model. If this model was validated, the model validator would enter information about that process. Similarly, a person responsible for model deployment would answer questions related to deployment and ongoing monitoring. If the lifecycle is not that structured, the person responsible for most of the work could fill in this template.

(6) **Evaluate Actual FactSheet with Consumers.** In this step an assessment is conducted of the quality and completeness of the actual FactSheet produced in the previous step. If the FactSheet is intended to be used by multiple roles (not uncommon), it should be evaluated separately for each role. To ground this assessment properly, each consumer should be asked to reflect and comment on how this FactSheet would actually help them perform their work or provide information to address their concerns.

(7) **Devise Other Templates for Other Audiences and Purposes.** This step returns to the beginning and is the only defined iteration in the methodology. There may be other consumers that need to be supported. Or, if the work so far has focused on the process of creating and deploying AI models, it would be worthwhile to consider consumers beyond this such as internal or external review boards or regulators, sales personnel, and end users or others affected by the product or service.
3 RELATED WORK

The drive to improve AI transparency via documentation has been motivated by the numerous high-profile examples of AI risk in society. This has sparked research groups [1, 3, 11, 13, 17, 18, 24] and government agencies [7, 19, 28] to explore what form this documentation should take. The main focus of these works has been on what should be included in this AI documentation for near universal use. Several of these efforts have included feedback loops, where society and industries have been able to comment on the proposals. Essential stakeholders for AI transparency are citizens who want to know more about the algorithmic decisions that are affecting them. This is challenging because the average citizen is not familiar with the technical details of AI. In recent work, Domagala and Spiro [9] engaged with citizens to better understand their needs regarding information about algorithmic systems deployed by the UK government. This was one way to approach Step 1 of the methodology described in Section 2. Earlier, Hind et al. [13] explored the needs of documentation producers via interviews and a documentation creation exercise lasting a few hours. This paper differs from these works since it examines AI documentation practice in situ through a months-long case study involving both producers and consumers.

Software documentation, in general, describes important characteristics of a system, application, or module. Despite its necessity, developers and software engineers find the task of documenting uninteresting and its creation often falls to technical writers who do not know all the details of the software they are documenting [20]. Not surprisingly, consumers of documentation generally dislike what is produced because such documentation is often incomplete, difficult to understand, or out of date [4]. But documentation is essential as it remains one of the few channels of communication between the developers of software and its various users [15].

For conventional software, researchers have developed sets of rules [20], evaluation criteria [22], and assessment methodologies [8, 23, 27]. Quality attributes such as accuracy, concreteness, writing style, and understandability have been offered as useful dimensions of quality [23]. Other dimensions such as completeness, unambiguity, conciseness, and ease of access have been proposed [20]. Yet other dimensions include consistency, traceability, reusability, format, trustworthiness, and retrievability [8, 30]. This paper draws on many of these dimensions to assess the quality of AI FactSheets.

4 EVALUATION METHODOLOGY

To answer our research questions, we partnered with an AI organization in the healthcare domain that was piloting the FactSheets methodology [25] for several models as a possible solution to their AI documentation and transparency needs. Data for the evaluation reported here consisted of two sets of interviews for each model: one for the FS team who followed the methodology to create the FactSheets, and one for FactSheet consumers to assess how the resulting FactSheets were meeting their needs. Together, these two perspectives provide a holistic view of the usefulness of the methodology.

4.1 Study Participants and Case Study Context

We divided participants into two roles: consumers and FactSheet (FS) team members. As in the FactSheets methodology, consumers are the intended end users of the created FactSheet. FactSheet team members are the ones driving the methodology for each of the models, defining the FactSheet template, and iteratively making the FactSheet. In total, we ran 17 interviews over 16 participants: seven FS team interviews and ten consumer interviews. One participant was

The FS team members were generally also the knowledgeable producers of facts in this case study. We ignore the distinction between FS team members and producers in what follows.
Table 1. The FS team (T\#) and consumer (C\#) participants and their roles for each model. The first character in each column header indicates the model under consideration, referred to as model A, model B, model C, and model D.

|       | A Role                  | B Role                     | C Role                  | D Role                  |
|-------|-------------------------|----------------------------|-------------------------|-------------------------|
| T1    | Data Scientist          | T3                         | Data Scientist          | T4                       |
| T2    | Lead Software Architect | T4                         | Data Scientist          |                         |
| C1    | Domain Expert           | T5                         | Data Scientist Manager  |                         |
| C2    | Product Manager         | C8                         | Consultant              |                         |
| C3    | Regulatory              | C9                         | Database Administrator  |                         |
| C4    | Customer                | C10                        | Customer                |                         |
| C5    | Business Analyst        |                            |                         |                         |
| C6    | Research                |                            |                         |                         |
| C7    | Data Scientist          |                            |                         |                         |

Table 2. The four models documented during the case study

| Model ID | Model Description                                                                 |
|----------|------------------------------------------------------------------------------------|
| A        | A set of models to assist Medicaid Fraud, Waste, and Abuse investigators in retrieving relevant information from Medical Insurance policy |
| B        | A mature model (version 21) to predict relative mortality risk of a hospitalization stay |
| C        | A model (in development) to identify potential risks of surgical complications to proactively mitigate these potential issues |
| D        | A model to identify populations that are likely to have improved health outcomes if their social factors are improved |

The participants in this case study are part of an AI organization responsible for creating and maintaining several different kinds of models for a variety of health and medical use cases. During the case study period, the team worked on documentation for four models (Table 2) and completed the FactSheets for models A and B. FactSheets for models C and D were in progress and had at least a first draft. Because of this, we were only able to interview consumers for models A and B. The FS team used the publicly available resources on the AI FactSheets 360 site [14] to help them follow the methodology. Resources included explanations of what FactSheets are and their intended applicability; a description of the methodology for creating FactSheets; example FactSheets; and links to research papers, videos and a Slack community.

4.2 Interview Protocols

For the FS team, we conducted approximately one-hour, semi-structured interviews that focused on how well the methodology worked or did not work in practice. More specifically, we focused on five main topics: (1) the context of, and motivations for, creating the FactSheet, (2) how they used the available resources, (3) how they implemented the FactSheets methodology within their actual context, (4) perceived benefits, and (5) perceived costs.
Interviews with the consumers lasted approximately 30 minutes. Unlike the FS team interviews, that focused mainly on the process of creating a FactSheet, consumer interviews focused on whether the created FactSheet met their needs, and if not, where additional or different content was needed. Topics in consumer interviews focused on: (1) evaluating the quality of the created FactSheet and (2) how the FactSheet has changed the way they work compared to the past.

Both FS team and consumer interviews were similarly run. In each interview, participants were given access to the finished (or in progress) FactSheet that they could refer to as needed. Interview sessions were run remotely, recorded, and transcribed for later analysis.

4.3 Data Analysis

To analyze participants’ responses, we used two approaches. For questions that expected more direct responses such as identifying missing content in the FactSheet, we identified and aggregated participants’ answers. For more complex questions such as the ones around perceived benefits and costs, we conducted thematic analysis [2], focusing on themes that addressed our research questions. First, participants’ responses were extracted from each interview transcript such that a complete response was available for each question, including any relevant context. One author extracted responses, annotating any references a participant made to the FactSheet as necessary.

All authors participated in the thematic analysis. The thematic analysis was an iterative and inductive process, with new themes emerging and collapsing as the authors worked through the data. Using the set of interview responses, each of the authors coded any sentences with a theme or topic relevant to the research or interview questions. Extracted codes and their definitions were iteratively added, consolidated, and removed as needed to form the codes. Upon completion, all three authors together discussed codes and consolidated codes into themes and subthemes. The emergent themes and subthemes contributed to our findings.

5 RESULTS

Our results are structured by research question. We address if the methodology was usable by a real-world FactSheets team not trained in human-centered practices (RQ1), if the resulting FactSheets met the needs of their various consumers (RQ2), and what the benefits and costs of this approach were seen to be (RQ3).

5.1 RQ1: Usability of Methodology by FS Team Members

Practitioners skilled in human-centered practices will recognize that the methodology applies these practices to the creation of FactSheet templates (and the resulting FactSheets). But is it reasonable to expect data scientists, engineers, and others not trained in these practices to execute the methodology and derive the expected benefits? In this section we report on how well FS team members were able to apply the methodology. Each subsection looks at a different aspect of the methodology’s usability: the value of available educational resources; how well individual steps of the methodology could be followed and how useful they were; and the ways in which the methodology helped in documenting AI models.

5.1.1 How FactSheet Teams Used Educational Resources. The FactSheets 360 website [14] provides an overview of the methodology, a number of illustrative FactSheets, and links to more detailed instructional content. When we asked the four FS team members who used the website to identify the resources that were the most useful to them, responses included the example FactSheets on the website (T1-A and T5-B) and the methodology summary (T2-A). When asked for more details, T1-A described how one example showed how disparate kinds of documentation could be brought together into a single place saying “It was self-contained. It provided all the info I needed to understand.” T5-B described
how the different views in the example FactSheets, such as the table view and slide view, helped him understand how FactSheets could be tailored to the needs of different consumers saying "Something clicked (for me) on one of the (example) pages about how the same facts can be represented in different ways, which was the key idea for me to really see the value of this approach. I haven’t thought about documentation in that way before". T6-D echoed this sentiment explaining how many of the discussions around what to document revolved around this perspective on documentation saying “There were two critical things on my mind. It was understanding the players. What do we mean by producers? What do we mean by consumers?” By framing documentation as something to be used by others, FS teams changed their orientation from just reporting on what a model did to focusing on how to make the description useful for specific consumers. More details on this observation follow in Section 5.1.2.

FS team members did not report that any resources were unhelpful. They did identify some gaps and stated that there was too much content to digest it all. T1-A identified procedural guidance that he would have liked to have such as a “minimum set of fields” required for a FactSheet template. T2-A described a desire to have concrete guidance for how long the methodology should take and which facts to start with. To address this, T2-A developed a schedule based on his team’s first attempt at creating a FactSheet template, which provided guidance for subsequent efforts. “We needed something to explain the process... Starting with the methodology of the site and then breaking it out into ‘Here’s a six-week schedule’ basically. To give (the rest of the team) something concrete.” Similarly, to address T1-A’s concern, the team shared the template (tailored through their work for the healthcare domain) and FactSheet for future teams to use as a starting point.

5.1.2 Methodology Steps. To better understand how well the individual steps of the methodology could be applied, we asked FS team members to reflect on the steps one by one. FS team members responded most often that Step 1 of the methodology, Know your FactSheet Consumers, was the most useful of the seven steps. Understanding their consumers encouraged FS teams to consider who they would be writing for, and what their documentation for them should include. Half the FS team members (T1-A, T2-A, T5-B) decided this step was the most important one. T1-A said, “Knowing your consumers was a step that was particularly useful for me... Once I had a particular consumer identified I realize that there are things that I need to think more about, or I need to reach out to people who know about this. So, it was useful.” Revealingly, T1-A first tried skipping Step 1 to save time, opting to go directly to getting feedback from consumers on a FactSheet based on an example in the FactSheets 360 website. He recalled, “We started filling it out without the right [consumer needs]... and then we kind of realized a lot of this stuff isn’t right. So we went back and rewrote quite a lot of it basically.” After this setback, the other FS team members decided to speak with potential consumers first.

Participants T2-A and T5-B noted the benefit of creating a first version of a FactSheet (Step 4) before gathering feedback from consumers on just the template (produced in Step 3). While creating this first version might be seen as adding unnecessary time, the conversations with consumers focused on making changes to this version instead of generating content from scratch. Since one of largest costs reported by participants in the FactSheets methodology was validating the FactSheet with consumers, FS teams appreciated this time-saver. T6-D noted the value of this step as well and added detail about how the act of filling out the FactSheet template spurred reflection on aspects of the model they previously had not considered saying, “It’s forcing us to answer questions about our model that we may have thought about, but never documented in any way.”

Although FS team members did not consider any specific steps of the methodology to be unhelpful, they did adapt the methodology to better fit their team’s needs and timeline. FS teams chose, for example, to minimize the seeming seriality of Steps 5 (have actual producers create a FactSheet) and 6 (evaluate the FactSheet with actual consumers).
Table 3. FS team member responses to if the methodology helped them in various ways. A response of ‘n/a’ indicates that the FS team member was unsure or unable to give a response.

| Did the methodology help... | T1-A | T2-A | T3-B | T4-B | T5-B | T4-C | T6-D | # Yes |
|-----------------------------|------|------|------|------|------|------|------|-------|
| ... Improve documentation consistency? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | 7 |
| ... Evaluate documentation usefulness? | Yes | n/a | Yes | Yes | Yes | Yes | Yes | 6 |
| ... Identify new documentation needs? | Yes | Yes | No | No | Yes | Yes | Yes | 5 |
| ... Facilitate creating useful documentation? | Yes | No | n/a | No | Yes | Yes | Yes | 4 |
| ... Improve documentation practices? | Yes | n/a | Yes | No | No | n/a | Yes | 3 |
| ... Identify new consumers? | Yes | Yes | No | No | No | n/a | No | 2 |

They preferred a more iterative approach where FactSheet content was evaluated as soon as there was enough for a specific consumer role, even if the rest of the information was not yet ready. These changes were implemented during the first pilot. Since the subsequent pilots were informed from the first, T3-B, T4-B, T5-B, T4-C and T6-D all followed the altered protocol, further compressing the timeline by starting with the template from model A.

5.1.3 How the Methodology Helped FactSheet Teams Document AI Models. FS team members noted how the methodology provided specific benefits over their previous documentation practices. Table 3 summarizes the questions about specific benefits that we asked about and participants’ responses. All FS team members agreed that the FactSheets methodology would likely enhance documentation consistency, primarily by consolidating what was previously scattered in multiple documents into a single place. T5-B summarized, “We’ll be able to consolidate in ways that make sense so that there’s one place for facts, rather than, lots of places to keep things up to date.”

T5-B and T6-D described how completing the FactSheet template had caused them to reflect on the usefulness of what they were writing in the FactSheet fields. T5-B described how a consumer need for making the problem description reusable encouraged conciseness. He said, “One of the best (consumer) feedbacks that we received was in the way we tried to condense the problem description in the FactSheet so that they could copy and paste it to engage with customers”.

Importantly, five of the FS team members agreed that the methodology helped their team identify new documentation needs. They discovered several additional template fields as a direct result of following the methodology. Table 4 shows most of the fields from the FactSheet template for model A. Newly-discovered fields are italicized. Examples included regulatory requirements, model maturity, usage considerations, and run time requirements. T6-D summarized the value stating, “There was a lot of information here that we wouldn’t perhaps otherwise made available or collected or even thought of.”

5.2 RQ2: FactSheet Usefulness for Consumers

To determine whether the FactSheets generated by the FS teams were useful to consumers, we asked consumers to (1) evaluate the FactSheet’s quality along several dimensions and describe any gaps and (2) to describe if and how the FactSheet differed from the AI documentation they had encountered previously.

5.2.1 Evaluation of Usefulness. We asked consumers to evaluate the FactSheets, along multiple quality dimensions, on a seven-point Likert scale ranging from ‘strongly disagree’ to ‘strongly agree’. The dimensions were drawn from work on assessing AI documentation quality [21] along with prior research on the pragmatics of effective communication. The dimensions and prompts are shown in Table 5. We found that the overall quality of the FactSheets produced was excellent, scoring high marks across all the quality dimensions as shown in Figure 1. These results suggest that the
work of the FS teams, perhaps especially the iterative evaluation of FactSheet content with consumers, paid off and resulted in highly relevant, useful, and usable documentation.

5.2.2 Things That Were Missing. Even with the high quality-dimension scores, consumers were able to identify things that were still missing. For any response to a prompt with a score of less than 'strongly agree', we asked the respondent what was missing or needed to be changed for that dimension. Table 6 summarizes their replies.

A closer look at this reveals that missing content tends to be quite role-specific. For example, C2-A’s suggestions are related to how the model fits into the larger business context. He wanted to see more information about which customers are using the model along with pointers to additional materials useful in discussions with potential customers. He stated, "It might be good to include something about where has the model been deployed?" The requests from C3-A, whose role involved regulatory concerns, focused on risk: the model context, risk evaluations, and risk mitigation. C4-A and C10-B, both in customer-facing roles, focused on how to determine model quality and how the quality compares to

| Table 4. The template created by the FS team for model A. Italicized fields indicate new fields that were added during the case study. |
|----------------------------------------------------------|
| **Overview**                                             |
| Model Name                                               | The name of the model                                   |
| Version and Dates                                        | The model’s version and date it was created             |
| Problem Solved                                           | A description of the problem the model is designed to solve |
| Maturity Ratings                                         | A evaluation of the model’s maturity along multiple dimensions |
| Contact Information                                      | Who to contact about this model                         |
| **Intended Use**                                         |
| Intended Domain                                          | The industry of functional area the model intends to support |
| Use Case                                                 | The use scenario, including who, what, when and how the model outputs intends to support in practice |
| **Data and Model**                                       |
| Training Data                                            | Statistics and details about training data              |
| Cohort / Therapeutic Area Definition                      | Population cohorts of interest for this use case        |
| Model Information                                        | Description of the model                                |
| Input/Output                                             | Description of expected model inputs and outputs        |
| Internal Test Data                                       | Description of the internal model test data used        |
| External Validation Data                                 | Description of the external model validation data       |
| **Performance**                                          |
| Performance Expectation & Robustness                     | Expected performance metrics and evaluation of the model’s adversarial robustness |
| Model Performance                                        | Description of model evaluation and actual performance metrics |
| Bias & Fairness                                          | Description and evaluation of the model’s fairness and bias |
| Explainability                                           | Description of the model’s ability to explain its predictions |
| **Usage Considerations**                                 |
| Optimal & Poor Conditions                                | Conditions under where the model performs well and poorly |
| Volume & Latency                                         | Expected number of predictions and time per prediction |
| Discontinue Use If                                       | Conditions under which to stop using the model          |
| Runtime Data Requirements                                | Data prerequisites necessary to run this model          |
| Runtime Technology Requirements                          | Technology prerequisites necessary to run this model    |
| Use-Case Tolerance of Error                              | Evaluation of the error tolerance of this model         |
| Market Differentiator                                    | Description of market differentiators for this model    |
| Related Areas Where Models May Be Useful                  | Other potential use cases for this model                |
Table 5. Likert-scale prompts given to consumers to assess FactSheet quality

| Quality Dimension | Prompt                                                                 |
|-------------------|------------------------------------------------------------------------|
| Completeness      | The FactSheet has all the information that I require for my use case.  |
| Evidence          | The FactSheet’s information is well supported with additional evidence provided where needed. |
| Vocabulary        | The FactSheet is written using appropriate vocabulary and word choice.  |
| Understandability | The FactSheet’s content and information is easy to understand.         |
| Layout            | The FactSheet’s structure and layout is intuitive.                     |
| Representation    | The FactSheet’s information is presented in the expected way with text, tables, and figures appropriately chosen. |
| Organization      | The FactSheet’s information was well organized and easy to locate.     |

Fig. 1. Consumer responses for evaluating the FactSheet quality. The responses on the right side of the dark vertical line represent responses of ‘slightly agree’ or higher.

competitors. C5-A, like C4-A, not having a deep data science background, wanted more guidance on how to interpret the model evaluation metrics, both in terms of how the model relates to the business, but also including the reasonable question of what makes a good metric score. P5-A said, “I think I understand what these (metric) scores might tell me, but I have to do a lot of thinking to connect it back to the business problem... What do they mean for me and the situation that I’m trying to evaluate? ... What’s a good range? What’s a bad range?” C7-A, a data scientist, wanted to see information that would help him improve the model, along with a history of model evaluations over time. These consumers’ role-specific requests underscore the point that different consumers have markedly different documentation needs, and provide further evidence that a one-size-fits-all solution for AI documentation will likely not suffice.

The combination of high quality scores for the FactSheet along with several suggestions for further improvement tell a somewhat mixed story. On the one hand, as indicated by the high scores, participants seemed to be satisfied with the content of the FactSheets. On the other hand, they still had suggestions for how to improve them. There are potentially several reasons for this. FS teams may not have iterated enough with consumers to gather all the relevant documentation needs (on average going through only two iterations). Or FS teams intentionally did not address all these needs to keep FactSheet size manageable (a role-based filtering mechanism, see 6.2, may help manage this trade off). Another interesting possibility is that by asking consumers in our interviews to reflect on a FactSheet from the perspective of the quality dimensions, gaps surfaced that they did not see before. If true, this latter interpretation may have implications for eliciting consumer needs more completely.
Table 6. Summary of missing documentation information per consumer. Italicized entries were mentioned by multiple consumers. Although market differentiation content existed in the FactSheets, participants wanted more detail than what was available, so it is included below.

| Consumer | Role                     | Missing documentation fields                                                                 |
|----------|--------------------------|------------------------------------------------------------------------------------------------|
| C1-A     | Domain Expert            | None                                                                                           |
| C2-A     | Product Manager          | Market differentiation; existing deployments; customers; pricing information; model maturity; links to marketing materials, presentations, and demos |
| C3-A     | Regulatory               | Market differentiation; Process model is replacing; benefits and costs of deployment; access information; model evaluation summary; known risks; applied risk mitigations; how to interpret model metrics; productization details |
| C4-A     | Customer-Facing          | Market differentiation; Mapping between model metrics to business use case; how to interpret model metrics; good/bad ranges for model metrics; business context |
| C5-A     | Business Analyst         | Good/bad ranges for model metrics                                                               |
| C6-A     | Research                 | Data sources; data cleaning details; data limitations and constraints; model training details; background of domain experts; definitions for bias; definitions for explainability |
| C7-A     | Data Scientist           | Existing deployments; opportunities for model improvement; customer feedback; history of model evaluations |
| C8-B     | Consultant               | None                                                                                           |
| C9-B     | Database Administrator   | None                                                                                           |
| C10-B    | Customer-Facing          | Customer-specific model evaluation                                                               |

5.3 RQ3: Methodology Benefits and Costs

To better understand the value proposition of both the methodology and FactSheets themselves, we asked FS team members and consumers what specific pros and cons they experienced, further reflecting on how this compared with their documentation experiences from before.

5.3.1 Methodology Benefits. As mentioned earlier, one key benefit of FactSheets was to consolidate disparate documentation. T1-A expressed how bringing in the different viewpoints gave a more holistic picture of the model and its context. He said, “When [you] have the document in front of [you] from all points of view, you start joining things together and asking questions... It gives a broader overall picture and fills gaps in your own knowledge.” This broader context would be difficult to capture without input from all the consumers. Another benefit of consolidation was how the FactSheet facilitated additional exploration. T1-A, T3-B and T6-D discussed how the more broadly written content of the FactSheet coupled with its pointers to where more details can be found enabled such exploration. T1-A said, “I think this is useful documentation for the reason that everyone can go in and find what they are looking for, maybe not at the level of detail that [they need], but there is links and hyperlinks to everything.” T6-D referred to the FactSheet as a “one-stop shop for the model”. Seven consumers (C2-A, C3-A, C4-A, C5-A, C6-A, C8-B and C9-B) agreed that the FactSheet acted as an effective point of entry for understanding the aspects of the model relevant to their role.

Consumers elaborated on the kinds of further exploration that the FactSheet enabled. For example, C2-A and C8-B explained how the FactSheet eased navigation to other relevant documents. C2-A said, “It’s a single document that makes it easy for me to navigate to the next level. (Previously), I’d have to hunt down multiple documents or ask people what was available... This seems to bring it together at least in a single launch point where I can read a bit of detail about it,
Table 7. The number of hours each FS team member spent working on the FactSheet

| T1-A  | T2-A  | T3-B  | T4-B  | T5-B  | T4-C  | T6-D  |
|-------|-------|-------|-------|-------|-------|-------|
| 15 hours | 24 hours | 10 hours | 8 hours | 15 hours | 24 hours | 6-8 hours |

and then I can understand where to go to find more information.” C3-A and C6-A explained that the FactSheet served as a pointer to the right people to contact for further follow up, potentially reducing the time spent gathering information about a model. C3-A summarized, “I think it reduces the amount of time that will be spent of clarifying issues... It would allow me to do pre-work without having to do 3 meetings with a group of 5 to 6 people.” C5-A and C7-A likewise agreed that it reduced the number of meetings required to understand the current state of a model. C1-A, C4-A and C10-B echoed similar sentiments for discussions with customers.

In addition to serving as a jumping-off point, FactSheets acting as the sole source (or at least sole anchor) of truth, enabled FS teams and consumers to reduce work in several ways. Previously, there would be several sources of documentation that sometimes provided conflicting information. The FactSheet as the single source of truth allowed content to be reliably copied for other creations such as customer-facing descriptions, reports, or presentations. C8-B recalled a time that she was able to reuse content in the FactSheet for a client. She said, “This is actually really nice that they’ve got this [documentation field] added in here because now I can lift this paragraph and put it into a client document... And in an approved definition or language for how to describe [it].” One FS team member, T3-B, likewise used the FactSheet as a text source for additional materials. Simply knowing where the most current documentation is also enabled more effective employee on boarding, especially when the people who worked on the model were no longer available. C6-A summarized, “You would talk to [someone] who’d go talk to someone else who has a link somewhere, or sends you some old weird spreadsheet that says, ‘go talk to this person’. And then someone would have left the company and then they don’t have access to something. So this is much much better, because it puts everything in one place.” Finally, FS team members also reported benefits centered around the theme of improved organization. They reported how FactSheets helped them get a better understanding of documentation needs earlier (T1-A, T2-A), reduce repetition (T2-A, T6-D), improve awareness of other documentation (T3-B), and help identify overlaps in existing documentation (T4-C, T5-B).

5.3.2 Methodology Costs. The most obvious cost faced by participants was simply the time required to create the template and FactSheet. FS team members reported times ranging from 6 to 24 hours (Table 7). This estimate includes all the steps of the methodology, including meetings with others and time spent filling out the fields. Of course, creating documentation of any sort takes time, and it is not clear that this is outside the norm. Scheduling time with consumers was the most common cost mentioned by FS team members, specifically, T1-A, T2-A, T5-B, and T6-D.

Aside from the challenge of getting people together, some steps of the methodology required additional effort from the FS team. One mentioned by FS team members T1-A, T3-B and T6-D was Step 4, filling in the FactSheet template for the first round of consumer feedback. This approach is meant to make the feedback sessions more streamlined as it asks consumers to critique something that exists instead of creating it on the spot. However, this approach passed the burden off to the FS team, leaving them to fill out fields they may be unfamiliar with.

Another challenge in preparing that first draft was locating some of the documentation that existed from past work on the model. Both T3-B and T6-D were working with a models that had scattered documentation. T6-D described how much of her time was spent just finding the source that had the information she needed for the FactSheet. She said,
“(The difficulty) is not knowing where to look, right? It’s reading through documents that you potentially didn’t have to read through.”

Other than the missing fields, noted above, consumers reported few specific costs to the methodology. The one exception was consumer C10-B who did not see the benefit of a FactSheet over the documentation that already existed for model B (with which she was already familiar). C10-B may be a bit of an outlier, however, as all nine of the other consumers found FactSheets to be better than their prior documentation.

6 DISCUSSION

The current case study looks at the early process of adopting the FactSheets methodology within an organization. It examines whether the methodology can be followed by those with no particular training in human-centered design, and whether the near-term benefits outweigh the costs. We believe this early picture is promising and expect even more gains from long-term adoption due to reuse of templates, growing awareness of what facts matter to consumers, and increased automation of fact collection.

6.1 Enabling Reuse

Despite our belief that there is no universal template or checklist for AI transparency, there is a strong and quite understandable desire to reuse an existing template, for example, one from [14], rather than developing one from scratch. What we have observed in the present work is that a core set of facts did apply in a very different domain from the ones considered to date. These included descriptions of model purpose, training data, and model inputs and outputs. As shown in Table 4, this core set was augmented with domain-specific facts, such as therapeutic-area definition (for these models in the healthcare space) and market differentiator for models that reached a level of maturity where they are competing for sales within a segment. This kind of “base plus extensions” form of reuse is likely to become a dominant form.

An additional observation from the present work is that once a template has been tailored for a particular organization, it can be used with very little additional work by others within that organization. While the models in this study were all in the healthcare space, they had very different types, use cases, and maturities. Even so, the template developed for the first model was readily reused for subsequent models.

6.2 Design Implications for Tools

One of the key gaps identified by FS team participants was the lack of tool support for gathering facts. Facts needed to be found in existing documentation (large, dispersed, and created for a variety of purposes for a variety of audiences), then adapted for use within the much more compact FactSheet, or were obtained by tracking down and talking with others in the organization. We believe this situation will rapidly improve in several ways:

- Fact automation: many kinds of facts can be captured automatically by suitably instrumented tools. Code repositories adapted for use in AI development can record key facts about training data and model versioning. Lightweight tools such as Python notebooks can include fact-capture mechanisms with very little overhead. As the field identifies the kinds of facts that best support transparency, this sort of automation will likely appear.
- Fact elicitation: some facts, such as model purpose or non-quantifiable aspects of potential bias, must be entered by a knowledgeable human. A good library of examples, wizards, and other elicitation tools will make this both more efficient and help make facts more consistent (supporting comparison) and consumable.
• Fact explanations: there will always be a wide range of expertise among the various consumers of FactSheets. Automatic mechanisms for tailoring fact content to skill level are not infeasible. In addition, a library of reusable links to definitions of terms and supporting material will allow consumers to better understand potentially complex aspects of AI without requiring fact producers to author this supplemental information repeatedly.
• Fact filtering: either through customized templates or filtering controls, consumers will eventually be able to customize documentation to their needs.

6.3 A Broader View of Transparency
A key focus of the research on AI transparency views it from the perspective of AI’s impacts on individual citizens and society. This is essential work. However, we must acknowledge that there are many stakeholders involved in the creation of facts about AI that collectively flow (or fail to flow) into documentation that may or may not be suitably transparent. AI transparency must grapple with this fairly untidy reality. The development of models is not a solitary activity. It involves cross-functional teams with different expertise, iteratively building, testing, refining, deploying, and monitoring AI models [22, 29]. Critically, these roles drop in and out at different points of the AI model development lifecycle [22]. Along the way, intermediate consumers of important facts about model development attempt to understand, and contribute to, the growing body of facts (the quality of their contributions being dependent on the quality of the facts they consume). Problematically, AI documentation is not designed to meet the needs of its various consumers. And while a single, high-level template may provide a level of basic transparency for some consumers, it will generally fail to generate the transparency needed by all.

We believe that documentation needs to be designed with a clear understanding of its various consumers. In the end, such documentation may have a multi-tiered structure with a near universal top-level view, supplemented by more detailed and specialized information in further tiers. This is a common approach for structuring complex information, for example, in areas such as product comparisons where the top view may include capacities and energy ratings with additional views providing detailed information on modes of operation, compatibility with other products, and multi-dimensional ratings. Our study suggests that at these more detailed levels, consumers’ documentation needs are going to be specific and unlikely to overlap across roles as shown in Table 6. This requires a human-centered approach to understanding and capturing these needs in templates and, eventually, supporting tools.

7 CONCLUSION
Our results suggest that the FactSheets methodology [25] worked as intended, that is, FS team members were able to successfully use the methodology to create AI documentation that was useful to its intended audience. Specifically, we found that

RQ1 Methodology usable without human-centered training? Even without training in human-centered design methodologies, FS team members adopted the methodology, successfully elicited needs from their consumers, and encoded those needs into a FactSheet template. Furthermore, the template addressed needs specific to the AI healthcare domain, but were still general enough to be reused across several, quite different, models.
RQ2 FactSheets addressed needs of consumers? Consumers found the resulting FactSheets met their needs, giving them high quality scores along several dimensions. Not all needs were met, however, and the identification of missing content further reinforced a core motivation for the FactSheets methodology: that AI documentation must be tailored to the specific needs of each consumer.
RQ3 Methodology benefits and costs? FS team members and consumers benefited from documentation that was authoritative, in a single location, and supported additional detailed exploration for specific needs. In 16 of 17 interviews, participants agreed that the FactSheets were an improvement over earlier documentation practices.

In summary, the benefits of following the FactSheets methodology to improve AI transparency seemed to outweigh the costs, and are best expressed by T2-A who closed their interview saying the following about the methodology, "I think it’s going to be very useful for us going forward and really it’s becoming a core part of what we’re doing."

REFERENCES

[1] M. Arnold, R. K. E. Bellamy, M. Hind, S. Hounde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity. *IBM Journal of Research & Development* 63, 4/5 (Sept. 2019).

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

[3] Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Joshua Joseph, Jessica Yurkofsky Keri Thomas, and Yue Chelsea Qiu. 2020. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. In *NeurIPS 2020 Workshop on Dataset Curation and Security*.

[4] Vikas S Chomal and Jatinender R Saini. 2015. Software Project Documentation-An Essence of Software Development. *International Journal of Advanced Networking and Applications* 6, 6 (2015), 2563.

[5] John Cook. 2021. Why the lying algorithms failed Zillow, and what it says about the business world’s love affair with AI. *GeekWire*.

[6] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

[7] Central Digital and Data Office. 2021. UK government publishes pioneering standard for algorithmic transparency. https://www.gov.uk/government/news/uk-government-publishes-pioneering-standard-for-algorithmic-transparency.

[8] Wei Ding, Peng Liang, Antony Tang, and Hans Van Vliet. 2014. Knowledge-based approaches in software documentation: A systematic literature review. *Information and Software Technology* 56, 6 (2014), 545–567.

[9] Natalia Domagala and Hannah Spiro. 2021. Engaging with the public about algorithmic transparency in the public sector. *Centre for Data Ethics and Innovation Blog* (June 2021).

[10] TM Forum. 2020. https://www.tmforum.org/resources/specification/ig1232-ai-model-data-sheet-specification-v1-0-0/.

[11] 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 (Dec. 2021), 86–92.

[12] Karen Hao. 2020. The UK exam debacle reminds us that algorithms can’t fix broken systems. *MIT Technology Review* (Aug. 2020).

[13] Michael Hind, Stephanie Hounde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R. Varshney. 2020. Experiences with Improving the Transparency of AI Models and Services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA ’20). Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383851

[14] IBM Research. [n. d.]. AI FactSheets 360 website. https://aifs360.mybluemix.net.

[15] Noela Jemutai Kipyegen and William FK Korir. 2013. Importance of software documentation. *International Journal of Computer Science Issues* (IJCSI) 10, 5 (2013), 223.

[16] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwiny. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. (2016).

[17] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *ACM CHI’20 Conference on Human Factors in Computing Systems*.

[18] 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 ACM Conference on Fairness, Accountability, and Transparency*. Atlanta, USA.

[19] Carl-Maria Mörch, Abhishek Gupta, and Brian L. Mishara. 2019. *Canada Protocol: an ethical checklist for the use of Artificial Intelligence in Suicide Prevention and Mental Health*. Technical Report.

[20] David Lorge Parnas. 2011. Precise documentation: The key to better software. In *The Future of Software Engineering*. Springer, 125–148.

[21] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ‘AI’ developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[22] Reinhold Plösch, Andreas Dautovic, and Matthias Saft. 2014. The value of software documentation quality. In *2014 14th International Conference on Advanced Networking and Applications* (ANA 2014), 6, 6 (2014), 2563.

[23] David Piorkowski, Daniel González, John Richards, and Stephanie Houde. 2020. Towards evaluating and eliciting high-quality documentation for “AI” developers. *IEEE Systems* 63, 4/5 (Sept. 2019).

[24] M. Arnold, R. K. E. Bellamy, M. Hind, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity. *IBM Journal of Research & Development* 63, 4/5 (Sept. 2019).
[25] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, Aleksandra Mojsilovic, and Kush R. Varshney. 2021. A Human-Centered Methodology for Creating AI FactSheets. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering (2021).

[26] Andrew D. Selbst. 2017. Disparate Impact in Big Data Policing. Georgia Law Review 52, 1 (Feb. 2017), 109–195.

[27] Karl L. Smart. 2002. Assessing quality documents. ACM Journal of Computer Documentation (JCD) 26, 3 (2002), 130–140.

[28] The European Commission’s High-Level Expert Group on Artificial Intelligence. 2020. Ethics Guidelines for Trustworthy AI. Brussels, Belgium.

[29] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–23.

[30] Junji Zhu, Vahid Garousi-Yusifoglu, Bo Sun, Golara Garousi, Shawn Shahnawaz, and Guenther Ruhe. 2015. Cost, benefits and quality of software development documentation: A systematic mapping. Journal of Systems and Software 99 (2015), 175–198.