Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle

Shipi Dhanorkar*
Pennsylvania State University
University Park, PA, USA
shipi@psu.edu

Christine T. Wolf†
Independent Researcher
Minneapolis, MN, USA
chris.wolf@gmail.com

Kun Qian†
Amazon.com, Inc.
Seattle, WA, USA
qiankq@amazon.com

Anbang Xu
IBM Research – Almaden
San Jose, CA, USA
anbangxu@us.ibm.com

Lucian Popa
IBM Research – Almaden
San Jose, CA, USA
lpopa@us.ibm.com

Yunyao Li
IBM Research – Almaden
San Jose, CA, USA
yunyaoli@us.ibm.com

ABSTRACT

The interpretability or explainability of AI systems (XAI) has been a topic gaining renewed attention in recent years across AI and HCI communities. Recent work has drawn attention to the emergent explainability requirements of *in situ*, applied projects, yet further exploratory work is needed to more fully understand this space. This paper investigates applied AI projects and reports on a qualitative interview study of individuals working on AI projects at a large technology and consulting company. Presenting an empirical understanding of the range of stakeholders in industrial AI projects, this paper also draws out the emergent explainability practices that arise as these projects unfold, highlighting the range of explanation audiences (who), as well as how their explainability needs evolve across the AI project lifecycle (when). We discuss the importance of adopting a sociotechnical lens in designing AI systems, noting how the "AI lifecycle" can serve as a design metaphor to further the XAI design field.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → *Enterprise computing*; • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

KEYWORDS

Explainable AI, Interviews, Work Practices

*Work done while research intern at IBM Research.
†Work done while working at IBM Research.

DIS '21, June 28-July 2, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8476-6/21/06…$15.00
https://doi.org/10.1145/3461778.3462131

ACM Reference Format:
Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle . In *Designing Interactive Systems Conference 2021 (DIS '21), June 28-July 2, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3461778.3462131

1 INTRODUCTION

From the subtle, predictive text suggestions offered by many text input interfaces, to the computer vision classifications that help homeowners identify intruders on home security camera footage or tag friends in snapshots, or the map-routing optimization apps that help shave minutes off of a commute, such encounters with artificial intelligence (AI) fill our daily lives. As AI applications become more widespread, the ways in which we interact with and come to understand these capabilities will shape not only how we relate to ourselves and to others, but also how we imagine and participate in our futures with artificially intelligent machines.

There are a number of challenges to the active and engaged participation of actors across algorithmic, data-driven ecosystems [53]. An essential challenge to confront is the problem of *awareness*, or the recognition that a sociotechnical system even includes AI [5, 15, 39]. Another challenge is comprehending how such capabilities work. This type of understanding shapes a person's ability to appropriately appraise the actions and outputs of AI models. Often signaled as *interpretability*, *explainability*, and *explainable AI (XAI)*, the human comprehension of AI systems is a topic that has gained wide traction in recent years within the AI [3, 11, 19] and human-centered design communities [1, 29, 51]. AI interpretability is not a new topic [9, 26]. The growing popularity of black-box modeling techniques lately has renewed attention to these important concerns. Black-box modelling techniques such as neural networks offer us incredible capabilities in the high accuracy with which they are able to perform various computational tasks. Yet despite this high accuracy, black-box models may learn patterns from training data that defy common sense, and in some domains, can cause harm. One cautionary tale is found in Caruana [8], which discusses

a neural network used to model hospital records and identify patients at high risk of hospital re-admission for pneumonia. The model learned that a patient's diagnosis of asthma signaled lower (rather than higher) medical risk. This defied clinical reality, however, since as a patient group, asthmatics need greater medical care and monitoring. Indeed, having an asthma diagnosis meant doctors treated pneumonia more aggressively and thus saw fewer hospital re-admissions (a conclusion that would be intuitive to any experienced medical professional). Missing from the model's worldview were other factors in the broader sociotechnical system, such as doctors applying their medical knowledge and treating asthma patients differently. This case illustrates why the powerful capabilities of AI models must be closely examined and aligned with the domain setting and work practices where the models' predictions are meant to be used. In other words, there must be domain experts "in the loop" to integrate data-driven technologies with domain knowledge and social reality. In the words of Ross et al. [42], humans need to feel confident that a model is "right for the right reasons."

As others have noted, there are ongoing needs to bridge the gap between AI algorithms and their experimental results with the actual settings of situated end use [52]. A wider aperture is needed to more fully understand AI explainability, one that considers the broader ecosystem of actors who hold stakes in an AI system [7, 13, 29]. Little is known, though, about how various actors in industrial AI projects come to understand and make sense of AI models, the material qualities explanations take, and the tensions that arise as situated actions unfold. Generative work is needed to understand practices of sensemaking and explaining and the real workplace needs and challenges actors face.

This paper investigates these concerns by reporting on a qualitative interview study of individuals working on industrial AI projects at a large technology company. This paper offers an empirical understanding of industrial AI projects, the range of stakeholders these projects involve, and the emergent explainability practices and concerns that arise as these projects unfold.
Across our participant interviews, we observe explainability concerns related to:

- Balancing external stakeholders needs with their AI knowledge
- Imbalances within teams internally
- Simplicity versus complexity tradeoff in designing explanations
- Concern about revealing too much

What do actors need to know about AI models? Our findings highlight a number of the different motivations that warrant explanations about AI models:

- Understanding its inner workings
- Details about data over which model is built
- Model design at a high level
- Ethical considerations baked in to the model
- Expectation mismatch
- Explanation in service of business actionability

We discuss related background work in the following section. Then we set out the methods and details of our qualitative interview study. We report the key themes in our findings, and we conclude the paper with a discussion of this work's implications.

2 RELATED WORK

2.1 Algorithmic advancements in Explainable AI (XAI)

Core concepts in XAI have been synthesized in various reviews [10, 19, 35]; a comprehensive review is outside the scope of this paper. At a high level, explanations are often described as being "local" (meaning the focus is the particular predictive output) or "global" (meaning the focus is the broader reasoning of the model overall). Many local explanation techniques output *post-hoc* explanations, meaning they provide information after the AI model is deployed. We find many different post-hoc techniques in the XAI literature. Much of this work relies on reduction and visualization techniques to provide interpretability that is compatible with human intuition. An explanation, for example, might provide text heatmaps [4, 28], (highlighting words/tokens in the input text) to interpret text-based models based on the underlying idea that different words have distinctive informative levels. Another approach might be to visualize the AI models, like t-SNE [49] which plots the model into a lower dimension scatterplot while preserving the structure of the original datapoints, or LIME, which uses surrogate models and sampling/weighting to make model approximations [41].

Another approach to XAI is during model design. AI models are often called either a "white box" or "black box"[43]. White-box models afford human comprehensability through logical expressions (e.g. rules, decision trees, logistic regression, etc). In black-box models, the relationship between inputs and outputs is obfuscated and often not easily intelligible to humans.

Even with these advancements in this research area, it is fraught with ambiguities that come with the association of many different terms to express the concept of "understanding" black box models. The literature has a constellation of interrelated and often interchangeably used terms: explainability, interpretability, transparency of models. According to authors in [18], explainability relates to being able to convey model outputs in ways that match human semantic concepts (interpretability) and simultaneously being complete in that description. Explanations are then needed to do two things: be able to summarise reasons for neural networks behavior while also providing an accurate representation of that summary. This offers exciting new material to imagine new forms of interactions and spur design innovations in this direction.

2.2 Designing for Human-centered AI

The explainability of AI is an inherently human-centered problem. Explanations of intelligent systems are not new. Research communities around expert systems [48] and recommender systems [20] have grappled with this concern before. Yet the renewed attention to explanations is a byproduct of two simultaneous processes: (i) growing realization of the substantial roles they play as arbiters in many facets of human lives and (ii) the acknowledgement of contemporary AI products' increased complexities and interdependencies. Miller articulates the need to ground XAI techniques with extant social scientific theories on how people explain, understand, and create shared meaning through interaction [33]. Similarly, recent work by Wang et al. [51] argues that XAI design can find inspiration in human reasoning processes, driving future social

science/AI synthesis. In particular, in their comprehensive network analysis of scholarly work in the area, Abdul et al. [1] note the need for further inquiry into real-world AI applications, which raise complex challenges around comprehensibility and understanding. The paradigm of human-centered explainable AI (HCXAI) invokes scholarship to be critically reflective of implicit assumptions and sensitive to the values and worldviews of various stakeholders [13].

Several studies have focused on building an empirical understanding of users' perceptions of different explanatory features. For example, Ehsan et al. [14] use automatically generated rationales in natural language to present the inner workings of a black-box model. Their analysis draws out the salience of contextual accuracy, awareness, and relatability in perceiving rationales as offering adequate justification and promoting understandability. A study of rationales in the form of decision sets (which are a bit distant in human-likeness as compared to natural language) revealed that rationales with newer definitions and longer mapping lists are associated with lower user satisfaction and higher response time. In considering how much of a model should be explained, one study cautions against the potential information overload engendered by a transparent approach in presenting model technicalities [38]. A large-scale experiment study on the effects of accuracy on trust in AI noted lay users overly trust a model when its observed accuracy is higher than their own accuracy [56]. This finding highlights the importance of establishing reasonable expectations and communicating inherent uncertainty in model predictions.

2.3 Designing Explanations using a socio-technical lens

Throughout the history of technology in society, various theoretical perspectives have profoundly contributed to unpacking the complexities between technological artefacts and the social contexts in which they are embedded. A common thread underlying situated action models [46, 47], distributed cognition [24], and activity theory [36] is the recognition how contexts inform the design of technological systems.

The contemporary AI landscape has been aptly characterized as opaque[6] with its ever-increasing interdependencies on multiple components and algorithmic sophistication. To bypass this opacity and to be able to thoroughly apprehend these technologies, Ehsan et al. [13] suggest studying the HCXAI paradigm through a socio-technical lens. This view acknowledges that technical systems cannot be abstracted away from the heterogeneous assemblages of social actors within which AI models are embedded. Accordingly, explainability needs unfold emergently in situated encounters between actors and AI [52]. AI models do not exist in a social vacuum. An AI system deployed in the real world has a wide range of stakeholders that extends beyond the immediate users (e.g., regulators, model developers, decision-makers, consumers) [21, 50]. Leveraging the perspectives of design practitioners, Liao et al. [29] uncover the different user needs that emerge from different user types. Specifically, researchers [55] showed how experienced designers take on machine learning (ML) as design material. With limited understanding of the underlying algorithms, designers rely on a combination of abstractions about the ML capability and exemplars as aids to engage collaboratively with data scientists and to design and develop novel interactions with ML. In addition to charting who is involved in an AI ecosystem, researchers must consider when those individuals become involved. Cai et al. [7] point to the human-AI onboarding phase as the time "when users are first being introduced to an AI system, learning its capabilities, and determining how they will partner with it in practice" (p. 2). We follow their attention to the temporalities of AI encounters. How might we think of human-AI interactions as situated in specific industrial cycles? Conceptualizing the "AI lifecycle" can serve as a socio-technical lens with which to explore the intersectionality of the social and technical components of AI encounters.

3 METHODS

We conducted an interview study to understand interpretability concerns that arise in industrial AI projects. We interviewed individuals employed at a large international technology and consulting corporation headquartered in North America (hereafter known as TechCorp, a pseudonym). We recruited informants via word-of-mouth, flyers posted on internal company message boards, and respondent-driven sampling [17].

Our empirical study focused on AI projects related to text data, a key application area within industrial AI [25]. Natural language is expressive. While language has an underlying grammatical structure, in use, it is rich, messy, and situated. These complexities make computationally modeling of language challenging. There is not only the intended meaning of the utterances that must be considered, but also context sensitivity, subtle markers, non-literal cues, and other linguistic devices (presenting social and emotional intelligence), to name a few.

To capture a variety of experiences and perspectives across text analytics projects, we recruited broadly and required each informant to have worked on a text project in some capacity and to have explained to another person how AI applications for text work (these explanations could have been delivered internally to project team members or externally to clients). Accordingly, our informants represent a range of roles across a variety of industrial AI-driven projects. Table 1 provides details about the participants included in our sample. In this context, there is overlap in the work practices of those in the researcher and data scientist roles, and the different titles reflect different organizational/reporting structures. While both roles leverage known/existing AI/ML algorithms and invent novel ones, researchers also work on client products/projects. It is important to note that our study does not include direct interviews with AI end-users or clients. Undeniably, including real users would have tendered increased ecological validity; yet their enrollment was practically constrained by their access (given the proprietary nature of TechCorp R&D projects). Second, targeting individual employees from TechCorp allowed us to take a focused approach to our subsequent data collection and analysis.

In total, we interviewed 30 individuals during August and September 2019. Interviews took place in person or via a web-based video call and typically lasted one hour each. We began by asking participants about their job roles and for a description of the AI/NLP projects from which they could draw AI explanation-oriented experiences. We asked them to reflect on these (including but not limited information about the stakeholders in the explanations, the

Table 1: Detailed description of the interview participants including their role, gender, and particular NLP project applications

ID	Job role	Gender	AI task and technology application domain
1	Researcher (AI/NLP)	M	Entity resolution/author disambiguation in medical publications, articles
2	Researcher (AI/NLP)	M	Debugging tools for AI/NLP researchers
3	Technical Strategist	M	Various (Q&A, knowledge base population, standardization)
4	Researcher (AI/NLP)	F	Classification of medical documents
5	Technical Strategist	M	Entity resolution in healthcare domain
6	Researcher (AI/NLP)	F	Analytics for service pricing
7	Data Scientist	F	Classification of statements in legal contract documents
8	Researcher (AI/NLP)	M	Document (RFPs, emails, chat messages) analysis for discovering actionable statements
9	Researcher (AI/NLP)	M	Topic modelling, embeddings for document similarity, search ranking, classification
10	Researcher (AI/NLP)	M	Conversational systems/human-in-the-loop planning
11	Data Scientist	M	Named entity recognition for financial application
12	Researcher (AI/NLP)	M	Text generation for HR application
13	Technical Strategist	M	Pharmacovigilance/drug safety
14	Researcher (AI/NLP) | Researcher (HCI)	M	Conversational quality of chatbots/human factors in AI
15	Data Scientist | Technical Strategist	F	Classification and clustering for skills inference and talent management
16	Product Manager	M	AI products for document analysis (e.g., legal contracts, invoices, purchase orders)
17	Data Scientist	M	Data mining (social media) - food safety and supply chain
18	Data Scientist	M	Sentiment analysis models for 10 languages (e.g., English, Korean, Hebrew)
19	Researcher (AI/NLP) | Researcher (Infoviz)	M	Gamification of AI |Visualization tools for black-box models
20	Researcher (AI/NLP)	F	NLP models for different domain and data types (e.g., finance, legal, call logs)
21	UX Designer	F	Human-in-the-loop tooling UI for domain experts in financial domain
22	Data Scientist	F	Customer complaints in call logs, customer reviews
23	Researcher (HCI/Infoviz)	M	Visualization tools for AI researchers
24	Researcher (HCI/Infoviz)	F	Sentiment analysis to predict employee engagement
25	Data Scientist	M	Dialog system designed to track and monitor supply chain infrastructure
26	Product Manager	M	Monitor performance of deployed AI
27	Researcher (HCI/ Information Viz)	M	Visualization tools for AI researchers
28	Product Manager | Technical Strategist	M	AI products for document analysis (e.g., legal contracts, invoices, purchase orders)
29	Researcher (AI/NLP)	M	Language model development for drug discovery
30	UX Designer	F	Interface design for AI products in financial domain

content that needed to be explained, the method by which the explanation was delivered, and the challenges that cropped up in the process of explaining AI models). We followed a semi-structured interview approach [37] that allowed informants to guide the discussion and contributed to an interactive exploration of our topic. During the interviews, many informants also showed us artifacts (via screenshare) that they had used to explain models (e.g., slide decks, scientific papers, dashboards, and infoviz apps). All interviews were audio-recorded and transcribed; we also took notes during these interviews. We analyzed the interview transcripts and notes following an iterative, thematic approach typical in affinity mapping [22].

4 FINDINGS

We frame our findings around the AI lifecycle, the broader complex process responsible for the development of any given AI product/service within an enterprise context. These findings emerged from participants' responses about their information needs: AI touchpoints, model development, model validation during proof-of-concept, deployment, and ongoing considerations with respect to explanations apparent throughout the AI lifecycle. Figure 1 summarizes our findings.

The issues surfaced by the participants are relevant to any client engagement, even though our informants focused on language processing systems. Some informants provided some examples specific to text processing by referencing particular NLP tasks related to entity classification, while others drew out concerns that were more overarching. We report on and elaborate these findings below.

4.1 AI touchpoints

4.1.1 Models' locations. AI models sit at different points within a sociotechnical system and a model's "location" shapes the types of explanations that are needed and for whom. Where the model sits within the ecosystem is also case-by-case and dependent on where that task fits within the overarching system/service. For example, in one project on which I-01 and I-05 collaborated, entity resolution was a central part of the service to pharmaceutical companies (identifying influencers in the medical domain). As such, the AI model sat "close" to end-use. In another project, however, a model handling NLP task of classification was further away from the end-use service (a platform that helped identify obligations in legal contracts). An AI/NLP researcher working on this project shared, *"We were working on a classification task, yes. But it is in service of some downstream task, not an end in and of itself. That's extremely important to remember, it's always for some downstream task"* (I-20).

4.1.2 Who interfaces with AI models? Our informants described how discussions about AI model technicalities take place amongst individuals with different backgrounds. A technical strategist remarked, *"Not all AI models are end-to-end. Some are directly human-facing and some are embedded in the enterprise pipeline. Depending on where they fit in, there could be different personas"* (I-28), listing trainer, developer, and end-user personas as those that are commonly referenced during client discussions. One approach that five of the informants (I-01, I-02, I-04, I-07, I-20, I-22) reported utilizing exposes what the black-box AI model has learned using linguistic expressions (i.e., forms of predicates or linguistic rules).

While these logical expressions provide comprehensible explanations to those researchers developing the model (and oftentimes the subject-matter experts working closely with them), they provide a far-too-fine-grained level of detail for other stakeholders. Researchers I-01 and I-12 brought up product teams (within the company) that do not drive technical aspects of the AI model during development yet are active participants in co-producing the desired AI product/service.

Our informants' responses highlight how even beyond TechCorp, an AI product/service might interface with different personas, resulting in different explanatory needs. Referring to a project in the financial sector, a project manager mentioned the different individuals who could be interacting with the AI tooling, noting, *"Initially, we had a machine learning operations kind of persona in the client data science team who is supposed to be relatively up to speed in what is involved in the caring and feeding of an AI model"* (I-26). Recounting the lack of such a definitive job title in subsequent engagements with the same client, I-26 mentioned that the responsibility for the model later fell on the IT operations team personnel, who had limited experience with AI models.

Other times AI product rollouts were accompanied by a more nuanced segment of model explanation-seekers. For example, the aforementioned product manager (I-26) mentioned *"[the] model developer (data scientist), model owner (who is in line with the business), [and] model validator (risk management personnel familiar with regulatory requirements and policies)"* as potential explanation-seekers.

We found particular notions of AI model explanations (e.g., those that are about how the model is built, why an output is manifested the way it is) appearing throughout the AI lifecycle and the related AI product/service within which the model was embedded.

4.2 Model building: conceptualizing and developing AI-based technical assets

An AI-driven service engagement begins with the development of initial model capabilities by model developers based on business needs and requirements. This initial stage entails rapid iterations and exploration to develop technical assets that serve as base models. Various scenarios prompt explanations at the development stage.

4.2.1 Understanding inner workings. For one, early in the AI lifecycle, model architects and developers primarily seek explanations to obtain a sense of why a model is predicting what it is and what linguistic features it might have learned. In describing an AI project on drug discovery, an AI/NLP researcher (I-29) echoed the natural curiosity that is inherent to researchers and how this trait influences researchers' desire to obtain a detailed explanation during development, saying, *"All scientists are skeptical people, they just don't want to be given an answer; they want to explore."* Other informants shared their own thoughts about models' inner workings. One researcher noted, *"Explanations can be useful for NLP researchers to explore and come up with hypotheses about why their models are working well"* (I-19). An AI/NLP researcher working on deep learning-based text generation AI models added, *"Low-level details like hyper-parameters would be discussed for debugging or brainstorming (amongst the research team)."* (I-12).

A recurring theme in our informants accounts was the notion of *actionability* as a motivation during model development and refinement efforts. During conceptualization and the initial development cycles (even prior to the release), explanations during *"model selection"* might give model trainers/developers a useful glimpse into the particularities mentioned a data scientist (I-18), as well as aid in *"comparing model results triggered by changing the underlying dataset or model drift,"* another data scientist (I-7) pointed out. Researchers often develop in-house or use existing explanatory visual tools to identify, track, and mitigate adverse model behaviors (e.g., inherent gender bias in which a model selectively associates the pronouns "he/him" with doctors and "she/hers" with nurses). Demoing a visual tool that he had developed to understand how a black-box language model worked for a specific domain, one researcher shared, *"If you know the layer and the head then you have all the information you need to remove its influence... by looking, you could say, oh this head at this layer is only causing adverse effects, kill it and retrain or you could tweak it perhaps in such a way to minimize bias"* (I-19).

4.2.2 Anticipating user questions. In the wings, researchers engage in deliberate and focused thought surrounding different questions: *"What XAI algorithms should I implement?...what questions might the user ask? And do I have the XAI technical capability to provide that explanation?"* are some of the questions raised by an HCI researcher (I-23). Another researcher (I-12) reported proactively anticipating user questions such as *"How have you generated these?"* in reference to an AI-driven HR application he had helped develop that auto-generated boilerplate language for job descriptions.

4.3 Model validation during proof-of-concept demonstration

Next, these initial AI-based technical assets serve as experimental proof-of-concepts, a base model which can then be "built on top of" and extended via customized learning for a particular client's industrial context. This "building on" approach requires explaining the contours of the particular base model, as one researcher (I-20) laid out. Such an explanation involves describing the data on which the base model was trained, what the model can do, and what the model cannot yet do, and then segueing into a discussion about the client's particular data, task, and business process/problem. Mapping the model's current functionality to the client's goals also involves discussing the resources (e.g., time, personnel, compute) the client is willing to devote to the project.

4.3.1 Details about data. An important part of discussions with clients revolves around data, and this was a dominant theme throughout informants' accounts. Discovery-motivated AI product offerings might involve incorporating novel sources of data (e.g., social media data) into an existing business process, which in turn requires explanatory articulation work. For example, one data scientist (I-17) spoke of a project working with a public health agency in monitoring food safety/food poisoning cases. In the existing process, food safety inspectors would review reports from calls to a food poisoning helpline, during which the operator would be able to collect detailed information from the person reporting illness. The AI project involved leveraging signals from social media posts to infer food poisoning cases. Although this allowed for a wider net

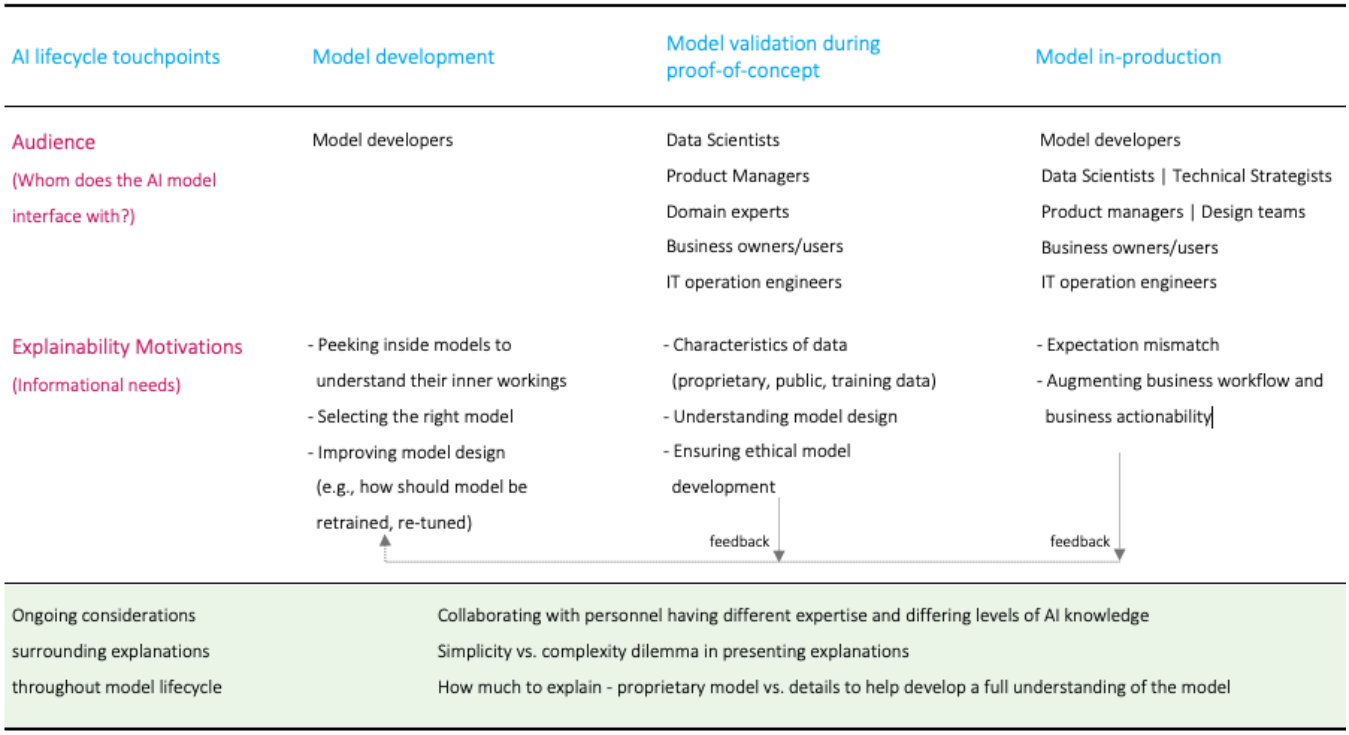

Figure 1: Audience and motivations for explainabilty as they appear in the AI lifecycle

(a person might not call a helpline, but they might post about their upset stomach on Twitter), it meant the data presented to the inspectors was sparser and noisier than they were used to. Inspectors had to cross-validate these with other data sources.

The data that are available (to train a model) and the data on which the client is hoping the model will be able to offer predictive outputs are key questions to clarify from the outset. For example, as in the case of a researcher who was working with academic scholars and doctors in a medical school on analyzing clinical documents. She shared: *"Domain experts want to know more about the public medical dataset that the model is trained on to gauge if it can adapt well to their proprietary data"* (I-4). Meanwhile, another data scientist shared, *"The client's data collection is not defined for the purposes of the ML project. My clients have done their core business, and now they want to reuse their data for new functionality in new and clever ways"* (I-22). But this data re-use is not self-evident, it requires careful dialogue with clients on goals, expectations, and resources. Often, clients will begin conversations with a speculative *"Well, what can the AI do?" to which an AI/NLP researcher described responding with, "Well, what kind of data do you have?"* (I-20).

4.3.2 Model mechanics at a high level. Bringing an AI product/service to life is a joint process between domain experts and model developers. However, information-sharing across a domain and model development undergirds this joint process. Subject-matter experts describe to model developers the technicalities of their domain; the onus of explaining the model's functioning and its behavior to the experts falls on the model developers. Our informants emphasized the shifting focus in the explanations that they offered. One AI/NLP researcher described, *"Initially we presented everything in the typical AI way (i.e., showing the diagram of the model). We even put equations but realized this will not work... After a few weeks, we started to only show examples and describing how the model works at a high level with examples"* (I-4). In this quotation we can see a recognition of the need to respect the expertise of each member of the collaborative service engagement.

Describing AI-based collaborations in the medical/life sciences domain, a technical strategist said, *"If we have to go in and teach doctors AI in order for these systems to work, we've lost the battle"* (I-03). He continued, explaining that domain users should be able to use the model's output in a context that is meaningful to their workflow. He noted how *"underneath [the model] the AI is happening and they don't necessarily need to understand how all that works. Someone needs to understand how it works and it should be explainable to the SME [subject-matter expert] if they need it. It's about appropriateness of the task"* (I-03).

4.3.3 Ensuring ethical considerations during model development. During the initial model development phase, clients and their legal teams provide high-level privacy requirements specific to their domain. These requirements feed into data pre-processing steps during which model developers generate masks and filters for personal data. Our informants discussed how in the proof-of-concept stage, model developers and associated product managers engage in transparent and open discussions about this data preparation, including the conceived mapping between privacy/security requirements and

data masks and filters to ensure appropriate legal compliance. Here, clients raise questions pertinent to regulatory issues with respect to the AI model. One product manager recalled a client asking, *"How do we know the model is safe to use? ... Users will ask questions about regulatory or compliance-related factors: Does the model identify this particular part of the GDPR law?"* (I-16).

4.4 AI product deployment, monitoring and maintenance

Explanations, of course, are also warranted when a model's behavior "in production" needs to be understood. Our informants' accounts are captured in two different themes: expectation mismatch, and the role of explanations in providing clients with actionable insights.

4.4.1 Expectation mismatch. An HCI researcher aptly characterized *"expectation mismatch"* as a common explainability scenario (I-27). An expectation mismatch is when the model's prediction differs from what the user expected the prediction to be. In the context of a client's business problem (e.g., coming up with a list of top 100 person recommendations who study non-small cell carcinoma in the European Union), one of the technical strategists described an AI service that uses entity resolution (a fundamental NLP task). The informant described this service as one that augments the current manual process of creating a list of key opinion leaders in the medical domain. Customers seek explanations when some names they expect to see on the list are not captured by the model. Referring to an erroneous name consolidation process, the informant explained, *"Data quality issues might arise resulting from expectation mismatch, but the list of recommendations must at least be as good as the manual process ... If they [the clients] come back with data quality issues ... we need to come back with an explanation of what happened"* (I-05). Another technical strategist indicated that explanations help in distinguishing outputs manifested through *"model mistakes versus decision disagreements"* (I-28). Still another technical strategist added, *"In human experience, when we are surprised, our expectations are violated. That is when we need some type of explanation. If we see or experience what we are expecting, then we just continue on without taking much note."* (I-3).

4.4.2 Explanations in service of business actionability. A dimension that shapes the appropriateness of an explanation is not only its level of technical detail but also its relation to some actionable item in the business realm. Like actionability during model development and improvement, actionability as it ties to business decisions was a key notion across informants. To distinguish, we label this "business actionability": the idea that explanations are provided in such a way that ties their insight to the broader business workflow within which they are provided. In our conversations with informants, we saw explanations as closely intertwined with business actionability, ultimately propelling a model's integration into the business workflow.

One data scientist (I-22) used the language of "mutable" and "immutable" variables to describe this concept. She brought up questions that client often ask her, *"Is the feature it is pointing out something I can change in my business? If not, how does knowing that help me?"*. Although the example she used to illustrate this point was not specifically related to text data (the number of dependents listed on a loan application), it demonstrates that explanations need to be tied to the overall business process if they are to provide meaningful guidance on what should happen next. The mutable/immutable concept also came up in other informants' accounts. For example, a product manager (I-26) described how it can be very difficult for users to understand compound variables, that models use in their decision making process because they do not translate to anything meaningful in the business domain. He went on to discuss the need to differentiate between explanations that imply some business action (i.e., are mutable in the real world). To generate actionable items, means injecting domain knowledge about which features are indeed combination-worthy and hang together. This could be accomplished via a pre-processing step, working with subject-matter experts and domain experts to understand what is indeed changeable and what is not for a given domain context.

4.5 Challenges and concerns in collaborations while engaging in explainability

Having explored the different forms of explanations through the lens of the AI lifecycle, we now turn to the various considerations and dilemmas our informants described. Our conversations surfaced some tacit, often overlooked sentiments surrounding AI-based collaborations across TechCorp.

4.5.1 External stakeholders and their informational needs. Informants recalled striving to provide the appropriate level of explanation to various stakeholders. Except for a few projects in which the AI product/service was designed specifically for technical users (e.g., tools for data scientists), there was a general sense among informants that there was a line beyond which explanations would be "too technical" for those who were not directly involved in the model development process. Furthermore, informants indicated that the line between what is "too technical" or "just technical enough" depended not only on the task and business case but also individual backgrounds and levels of technical AI literacies of the team. Describing a use case to leverage AI techniques for document retrieval, an NLP researcher noted that *"It always requires some creativity to explain technical things to non-technical people"* (I-09).

What constitutes an explanation is highly contextual and situated in the AI/ML knowledge that stakeholders bring. According to one designer, *"We have to balance the information you are sharing about the AI underpinnings, it can overwhelm the user ... They aren't dumb; it's not that at all. They are just looking at it from a different perspective, from a business perspective. So they are only concerned with numbers or results as they relate to the business problem"* (I-21). One AI/NLP researcher drew upon his experience in providing explanations: *"Depending on who is at table, explanation-oriented discussions could be only about precision/recall numbers"* (I-8). In contrast, a product manager mentioned how explanations might be better delivered without performance numbers because *"numbers can be a point of discomfort"* (I-16).

We also observed the deliberate effort that model researchers need to make to debunk the myth that AI is magic. A data scientist mentioned that to some clients, *"[AI] may seem like a magic box that can do anything"* (I-22). An AI/NLP researcher expressed frustration with this notion of AI, remarking, *"The whole idea that this is magic is a sad by-product of all the hype ... None of this is magic. It's just*

statistics. Deep learning is all derivatives; that's calculus. Isaac Newton says hi! But it's hard to get people to understand that because if you don't understand something, it seems like magic. But then if it doesn't work well, you need to understand why" (I-20).

4.5.2 Concerns within team personnel. In addition to the explanatory articulation work that happens with clients, our interviews also evoked nuanced thoughts about how the informational needs of internal team members sometimes are purposefully left unmet. One AI/NLP researcher said, *"High-level concepts (what types of information the model considers and how it bootstraps off information organization schemas like MeSH*) is sufficient for product teams, along with accuracy"* (I-01).

Offering a more nuanced reaction, one product manager (I-16) brought up AI's *"brutal, steep learning curve"* that he had to overcome to be able to intelligibly converse with researchers and data scientists with whom he was working. According to one researcher, such knowledge gaps in team personnel elevate tensions between AI/ML researchers and other team members who might feel a *"loss of control,"* making the *"support and maintenance of explainable models hard"* (I-08).

Reflecting on an ongoing project, a designer working on the interface of a human-in-the-loop tool to support classification of entities in documents (I-30) described the struggle to understand the linguistics and the disconnect she experienced from the AI model for which she was designing. I-30 shared her screen to demonstrate the AI tooling to which she was referring. *"I don't know how that works,"* I-30 said and described her uncertainty about which elements the design should incorporate.

4.5.3 Simplicity versus complexity. A central trade-off raised in many interviews that is very important for XAI is the idea of *"simplicity versus complexity,"* as one researcher (I-27) put it. He described how *"the design space for (explaining) models to end users is in a way more constrained than for expert users. In certain respects, [it is] the hardest design challenge that you take on because you have to assume that you have to put in a very, very very shallow learning curve. It cannot be as steep as it would be for an expert"* (I-27). Describing a gamification project to explain black-box models, another AI/NLP researcher echoed this necessary focus on clear explanations, remarking, *"It is hard to communicate without using words like 'activation' or 'clustering'"* (I-19).

Explanations have been manifested using visualizations in the XAI literature. Different AI expertise levels, however, shadow their usage in real-world projects. One product manager described how a very well-known explainability technique (LIME [41]) was not readily usable with business users, saying, *"In terms of how they are visualized, originally [we] just had a graphic representation of LIME with a percent and the feedback we got back was that was too complicated. Business users found that confusing, didn't add up to the confidence rating/interval. So they asked for just a text summary"* (I-26).

4.5.4 How much to explain. Another often overlooked, practical concern informants raised is how much to explain and the contours surrounding that explanation. On one hand, for AI models to be truly incorporated into workflows, it is imperative that users develop a comprehensive understanding of them. On the other hand, there are concerns about disclosing too much about proprietary AI technology. According to one technical strategist, *"We never sit down and explain"* (I-05). An AI/NLP researcher remarked, *"We never get into the weeds (with clients)"* (I-8). Referring to a project on building sentiment models, a data scientist noted, *"[The] developer has to provide all these constraints but also needs to balance without revealing too much about how it [the model] actually works"* (I-18).

Furthermore, explanations can lead to unwarranted situations if they reveal too much. *"Explanatory features can reveal identities (e.g., easily inferring employee, department, etc.),"* mentioned I-24, an HCI researcher, while recounting one project analyzing a workplace social forum to predict employee engagement in an organization.

5 DISCUSSION

Human comprehensibility of AI has been a topic of renewed attention and investigation in recent years as the use of black-box models becomes more common. This paper contributes to these conversations by offering an in-depth, qualitative view of industrial AI projects and the *in situ* sensemaking practices that unfold in these projects. Our approach offers us a set of complementary points-of-view on the topic of interpretability and explainabilty of AI models. Our inquiry answers the calls set out in [1, 13, 52], filling important gaps around our understanding of real-world explainability concerns. At the same time our work raises questions in need of further investigation.

As the accounts of our informants reveal, real-world industrial AI projects involve diverse collectives of actors. Each person brings to the team their own technical literacy/background and comfort level with the technical underpinnings of complex technologies like AI. They each, also, have their point-of-view and motivating ambitions vis-à-vis the project/service under development. For researchers and data scientists, an AI text project represents a challenging research problem; for product managers and technical strategists, it represents an opportunity to shape industries and deliver cutting-edge services; and for designers, it represents a complex translation problem from user needs to design elements. Although we have not interviewed the customers our informants serve, from these accounts we are able to gain a view that their interactions with customers are collaborative and framed by the broader services relationships with which they emerge; such insights provide further support for the need to think of explanations as iterative, interactive, and emergent, rather than a static quality of model (explainable or not).

This paper brings attention to explainability practices in industrial AI projects, specifically practices in relation to AI model touchpoints, the multitude of ways in which explanations become warranted, the *in situ* nature of the explanations, how explanations of models aid in resolving expectation violations, and how actionability is closely tied to models' explanations. This analysis also acquaints us with the different explainability challenges and constraints that dwell within an entanglement like the AI lifecycle.

5.1 The AI lifecycle as a lens for XAI design

XAI research within the technical AI community focuses on a host of algorithmic techniques, but rarely addresses explanations in the sense of envisioning them as occurring at different AI touchpoints.

A narrow focus such as this is likely to lead to designs that are self-consistent in functionality, but fail in practice because they do not consider the *how* or the *why* of situated explainability practices.

In this paper, we have examined how the explanations our informants described were situated in specific stages of the AI lifecycle. Explainability needs during model development differ from those that emerge during model validation, model deployment and so forth. Our informants' accounts cast in sharp relief the different explanations sought during these stages - debugging and brainstorming during initial model building; data, high-level model understanding, and ethical considerations during model validation; expectation mismatches and business actionability during AI deployment.

Our use of the "lifecycle" metaphor signals an organic and situated ecosystem. Using this as a design metaphor invites us to trace how actors and concerns circulate across the AI lifecycle - such tracings help sketch the range of actors for whom explanations might arise or be owed and how those concerns move and morph along the lifecycle. Further, in mapping out the AI lifecycle and where and how different actors reside within it, assumptions may also be interrogated to reveal underlying values and norms on who and what is made (in)visible in our metaphors of complex systems like AI. Who holds a stake in understanding this AI model's outputs? Where do they come to interface with that model, along the AI lifecycle? How do those interactions change over the course of a project's lifespan?

Using the AI lifecycle as a design metaphor for the XAI realm naturally evokes the cycles involved in software engineering (SE). In the SE development lifecycle, the focus is on building software that is primarily deterministic in nature. In contrast, AI and ML models are inherently probabilistic. Therefore, uncertainty lurks in the AI lifecycle. But this uncertainty is not always a negative element, it can be generative for design, as recent work by Benjamin et al. [27] explores. That work introduces a conceptual vocabulary around AI uncertainty to tease apart ways in which that uncertainty might serve as a design material. One salient point from this work is to note that uncertainty is just "part and parcel" of how AI systems work: a predictive model's raison d'etre is taking action (i.e., assigning a predictive label) amongst uncertainty (i.e., real-world data inputs). Making plain this inherent uncertainty then raises design questions around if, how, and when other actors come to encounter this uncertainty - and how those encounters may be scaffolded through design.

In adopting a lifecycle view on AI, we must remember that AI models are actors in the sociotechnical sense. This means models have forms of material agency and those agencies, along with others, play a role in shaping interactions [27, 32, 40]. We can see this in forms of "machine teaching", where humans interact with and "teach" the model by providing feedback on data examples the model offers [32]. But material agency can also be a provocative design prompt, challenging our tacit assumptions of human primacy in sociotechnical systems. For example, Reddy et al. [40] propose design methods that speculate: what if an AI model had an ethical agenda? Applying ethical and moral valence to machine action in this way forces us to confront our entangled and precarious position as ethical actors - together with nonhuman actors - in complex sociotechnical ecosystems.

5.2 The relationship between local explanations and a preceding, shared understanding upon which to build

Overwhelmingly, the XAI literature focuses on providing local explanations – that is, explanations about a specific input/output. Why did the model take this action given this input? A focus on the local aligns with many of the sensemaking practices we uncovered during our empirical study. But our study also reveals the relation of individual, instance-level explanations to higher-level understandings of the AI system and lifecycle. Informants talked of providing a "high level" overview, which offers a more macro- and meta-level explanation of the AI model's technical mechanics. This high-level overview led to a shared understanding of a model's capabilities in relation to the data at hand and possible avenues for model enrichments or customizations. This shared understanding was achieved by framing high-level explanations, with specific customer business use cases. What are we doing here? How is this going to support your business? The business context is the specific "landscape" upon which every industrial AI project takes place; this industrial landscape frames what explanations are needed and the actionability they ought to signal to users.

In many of our informants' accounts there is a temporality and sequencing – a high level overview helps to create a shared understanding at the start of an endeavor, negotiates reasonable expectations on outcomes and performance, and lets the AI development work proceed. Then as the work progresses, and AI actions are seen on individual outputs, the need for further detailed explanations arise. Such temporal ordering – and the common grounding it builds upon – offers an avenue for future work to explore. How can shared understandings amongst stakeholders in AI model ecosystems be seeded? Yang et al. [54] provide useful insights here, noting how various AI/ML topics and themes can be used to spark fruitful design dialogues. We must also take into account individuals' AI literacy [30]. How can AI design practice accommodate ecosystems with diverse AI literacies as AI development work progresses? These are open questions.

5.3 Expanding the ambit of explanations

Characterizing AI models as "explainable" or not is a simplistic view of what it means to understand AI. Human comprehension of these complex systems emerges from the various conversations between model builders/product managers and business clients; through these situated and unfolding dialogues, parties develop a shared understanding about a particular AI project [23]. As we have seen, explanations are not "one-size-fits-all", and their form and function depend on the context within which explainability needs emerge.

Our empirical findings surfaced efforts to demystify the "magic" some associate with AI. For stakeholders who are removed from the nitty-gritty of AI model development, equating AI with magic implies that its capabilities are spectacular, far too impressive, and cannot be subject to examination. Science and technology studies (STS) literature has long looked at the ways in which humans attribute technology with magic [45]. As the Arthur C. Clark quote goes: "Any sufficiently advanced technology is indistinguishable from magic." But, as Gell [16] notes, an ascription of magical prowess to

technology makes it seem as though technological development is *costless* simply because its technical details are hidden from view or often made irrelevant to the outputs it produces. Rendering technological cost invisible can be dangerous, especially in assessing a technology's attendant risk. Risk is a socially-constructed concept, shaped by cultural values and norms, as Luusua and Ylipulli [31] note: "risk is a design decision like any other." The risks around AI must be made visible in the design process - doing so not only raises awareness around risk (moving beyond the view of AI as costless magic), it also opens up risk as a site of possible design intervention. How might we design the risk around the technological artifacts and services we are building? Dove and Fayard [12] offer insight here - exploring the metaphor of monsters and monstrosity in the AI design process. Making visible and material the socio-cultural values, assumptions, and beliefs around AI enacts a richer, messier design practice, capable of wrestling with- and trying to intervene-in our monstrous sociotechnical ecosystems.

We make note, though, of trade-offs in trying to break through AI's black box. In our study, we uncovered emergent and opportunistic strategies used by researchers to make AI accessible to collaborating partners (e.g., showing the client the model diagram with equations in the beginning and then changing course to show the client examples of how the model behaves). These strategies are suggestive of a perplexing dilemma that confronts the project team: coming up with explanations suitable for domain experts and business clients, while also building an overall mental model for clients (understandability) that does not leave out crucial information about the model (completeness). Very recent work in this space has suggested different visualization techniques that can be customized to a person's desired level of detail and cognitive load [2, 57]. Allowing for users to tweak explanations to specific needs and instances while taking into account cognitive load considerations can help users form accurate, yet flexible mental models.

This paper extends this prior work by bringing these issues in conversation with a sociotechnical perspective. Our work offers emergent considerations to further the XAI design space: (i) enriching ongoing dialectical exchanges between stakeholders by using rich artifacts (e.g., Model Cards [34]); (ii) seeking alignment with a given model's location in the AI lifecycle (e.g., embedding the progressive disclosure principles [44]); (iii) catering to wide variations in stakeholders' AI literacies; (iv) minimizing expectation violations during deployment (e.g., leveraging pseudo learning-by-doing); and (v) balancing technical details with high-level understandings as appropriate to the domain and task at hand.

5.4 Designing for service touchpoints

As we have seen, explanations are tied closely to actionability. What is this information telling me, what can and what ought I do with it? For AI engineers and researchers, explanations are useful in debugging and probing model behavior. On the other hand, for customers touching AI models at a different part in the AI lifecycle, explanations are needed to help create an understanding of the model's limitations and boundaries (the training data's point-of-view) and the impact it would have on the business context. Rather than focusing myopically on performance measures alone, it's about matching the capabilities of AI to business needs and thoughtfully embedding those capabilities into human-in-the-loop processes to ensure key business objectives are met. In our informants' accounts, we can see how matching AI capabilities to situated business settings is an ongoing, collaborative process between a number of stakeholders. This echoes recent work by Hong et al. [23], noting how interpretability involves cooperation and trust-building among stakeholders. AI models in real-world industrial projects are encountered within a services frame; they are embedded within a broader sociotechnical service system and are not artifacts taken "off the shelf." Models are not "plug and play" but instead a site of ongoing sensemaking and collaborative learning amongst stakeholders. AI engineers must continually learn and stay "up to date" on the model as changes and tweaks are made; product managers must gain a fluency in algorithmic techniques to lead the team strategically; designers and strategists must iteratively design and re-design interactive experiences. In pointing to the dynamic, lively encounters amongst these stakeholders – and the shared understanding they ongoingly strive to create and maintain – we hope this work offers insights for future research into the collaborative and situated dimensions of AI sensemaking and explainability.

6 LIMITATIONS

This empirical study with informants working on AI projects offers a unique vantage point from which to understand how explanations are constructed and shaped in real-world projects. Here, too, we qualify the findings of our study and acknowledge our limitations. First, in this study, explainable AI was characterized specifically, using text-based projects. Explainability practices in other data realms might differ. While we were able to capture the views of corporate research and product teams (e.g., HCI researchers, data scientists, designers, technical strategists, and product managers), our research does not directly reflect the views of client end-users. Capturing client end-users' accounts and opinions around explainability concerns is an important part of the equation here, but our access did not allow for interviewing people in these types of roles. Despite this limitation, our in-depth interviews with individuals in a range of professional, corporate roles in industrial AI projects make our findings quite rich and offers (if obliquely) a view to client end-users' concerns, as they are experienced and recollected by project members.

As mentioned before, all informants were working on projects at a single corporation, TechCorp. Specific organizational processes and organizational cultures no doubt shape explainability practices in industry. We anticipated this problem and to help address it, we made deliberate efforts to recruit participants across TechCorp organizational departments (thus working on different product offerings). Additional work is needed to more fully understand the organizational dimensions of explanation practices, which we hope this paper meaningfully motivates future human-centered XAI research.

7 SUMMARY AND CONCLUSIONS

In this paper, we have taken up the topic of explainable AI in real-world enterprise projects. Our study provides an in-depth, qualitative view of text analytics projects at a large technology corporation and how sensemaking practices unfold in these projects. We have

described and illustrated these practices with examples from informants about how and when explanations arise. Our findings highlight the nuanced efforts at model explanations that go beyond algorithmic artifacts like performance numbers. In doing so, our study broadens the typical XAI-focused gaze to consider not only AI models in and of themselves, but also with whom the models might be interfacing, at what location along the AI lifecycle, and for what purpose. This, of course, only begins to chart the complex, sociotechnical ecosystems that we find around AI models. Ongoing cartographies are needed to further deepen our AI design praxis as it continuously unfolds in contemporary life.

ACKNOWLEDGMENTS

We would like to sincerely thank our participants for their time and their insights. Thank you to the reviewers who helped strengthen this work. This work was completed when the first author was a research intern at IBM Research. All opinions expressed are our own and do not reflect any institutional endorsement.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3173574.3174156

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[4] Malika Aubakirova and Mohit Bansal. 2016. Interpreting Neural Networks to Improve Politeness Comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2035–2041. https://doi.org/10.18653/v1/D16-1216

[5] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (2017), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

[6] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512. https://doi.org/10.1177/2053951715622512

[7] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. https://doi.org/10.1145/3359206

[8] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[9] Mark G. Core, H. Chad Lane, Michael van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. 2006. Building Explainable Artificial Intelligence Systems. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence - Volume 2* (Boston, Massachusetts) *(IAAI'06)*. AAAI Press, 1766–1773.

[10] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 447–459.

[11] Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2020. Tutorial on Explainability for Natural Language Processing. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (2020).

[12] Graham Dove and Anne-Laure Fayard. 2020. Monsters, Metaphors, and Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3313831.3376275

[13] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 449–466.

[14] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 263–274. https://doi.org/10.1145/3301275.3302316

[15] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300724

[16] Alfred Gell. 1988. Technology and magic. *Anthropology Today* 4, 2 (1988), 6–9.

[17] Krista J. Gile and Mark S. Handcock. 2010. 7. Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology* 40, 1 (May 2010), 285–327. https://doi.org/10.1111/j.1467-9531.2010.01223.x

[18] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[19] 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.

[20] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/358916.358995

[21] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 16–19.

[22] Karen Holtzblatt, Jessamyn Wendell, and Shelley Wood. 2004. *Rapid Contextual Design*. Morgan Kaufmann. https://www.elsevier.com/books/rapid-contextual-design/holtzblatt/978-0-12-354051-5

[23] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 068 (May 2020), 26 pages. https://doi.org/10.1145/3392878

[24] Edwin Hutchins. 1995. *Cognition in the Wild*. Number 1995. MIT press.

[25] Ashwin Ittoo, Le Minh Nguyen, and Antal van den Bosch. 2016. Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry* 78 (May 2016), 96–107. https://doi.org/10.1016/j.compind.2015.12.001

[26] W. Lewis Johnson. 1994. Agents That Learn to Explain Themselves. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence* (Seattle, Washington) *(AAAI'94)*. AAAI Press, 1257–1263.

[27] Jesse Josua Benjamin, Arne Berger, Nick Merrill, and James Pierce. 2021. Machine Learning Uncertainty as a Design Material:A Post-Phenomenological Inquiry. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14.

[28] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 681–691. https://doi.org/10.18653/v1/N16-1082

[29] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[30] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[31] Aale Luusua and Johanna Ylipulli. 2020. Artificial Intelligence and Risk in Design. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1235–1244. https://doi.org/10.1145/3357236.3395491

[32] Nirav Malsattar, Tomo Kihara, and Elisa Giaccardi. 2019. Designing and Prototyping from the Perspective of AI in the Wild. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 1083–1088. https:

//doi.org/10.1145/3322276.3322351
[33] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007
[34] 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
[35] 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
[36] Bonnie A Nardi. 1996. *Context and Consciousness: Activity Theory and Human-Computer Interaction.* MIT Press.
[37] Judith S. Olson and Wendy A. Kellogg. 2014. *Ways of Knowing in HCI.* Springer, New York, NY, New York, NY, USA. https://doi.org/10.1007/978-1-4939-0378-8
[38] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–52. https://doi.org/10.1145/3411764.3445315
[39] Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/2702123.2702174
[40] Anuradha Reddy, Iohanna Nicenboim, James Pierce, and Elisa Giaccardi. 2020. Encountering ethics through design: a workshop with nonhuman participants. *AI & SOCIETY* (2020), 1–9. https://doi.org/10.1007/s00146-020-01088-7
[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778
[42] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2662–2670. https://doi.org/10.24963/ijcai.2017/371
[43] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x
[44] Aaron Springer and Steve Whittaker. 2019. Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 107–120. https://doi.org/10.1145/3301275.3302322
[45] William A Stahl. 1995. Venerating the black box: Magic in media discourse on technology. *Science, Technology, & Human Values* 20, 2 (1995), 234–258.
[46] Lucy A. Suchman. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication.* Cambridge University Press, USA.
[47] Lucy A Suchman. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions.* Cambridge University Press.
[48] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325.
[49] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html
[50] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014
[51] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831
[52] Christine T. Wolf. 2019. Explainability Scenarios: Towards Scenario-Based XAI Design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 252–257. https://doi.org/10.1145/3301275.3302317
[53] Christine T. Wolf, Haiyi Zhu, Julia Bullard, Min Kyung Lee, and Jed R. Brubaker. 2018. The Changing Contours of "Participation" in Data-Driven, Algorithmic Ecosystems: Challenges, Tactics, and an Agenda. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. Association for Computing Machinery, New York, NY, USA, 377–384. https://doi.org/10.1145/3272973.3273005
[54] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173704
[55] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 585–596. https://doi.org/10.1145/3196709.3196730
[56] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509
[57] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 1245–1257. https://doi.org/10.1145/3357236.3395528