1 A Human-Centered Agenda for Intelligible Machine Learning

Jennifer Wortman Vaughan and Hanna Wallach

Abstract. To build machine learning systems that are reliable, trustworthy, and fair, we must be able to provide relevant stakeholders with an understanding of how these systems work. Yet what makes a system "intelligible" is difficult to pin down. Intelligibility is a fundamentally human-centered concept that lacks a one-size-fits-all solution. Although many intelligibility techniques have been proposed in the machine learning literature, there are many more open questions about how best to provide stakeholders with the information they need to achieve their desired goals. In this chapter, we begin with an overview of the intelligible machine learning landscape and give several examples of the diverse ways in which needs for intelligibility can arise. We provide an overview of the techniques for achieving intelligibility that have been proposed in the machine learning literature. We discuss the importance of taking a human-centered strategy when designing intelligibility techniques or when verifying that these techniques achieve their intended goals. We also argue that the notion of intelligibility should be expanded beyond machine learning models to other components of machine learning systems, such as datasets and performance metrics. Finally, we emphasize the necessity of tight integration between the machine learning and human–computer interaction communities.

1.1 The Intelligible Machine Learning Landscape

People are at the heart of each and every stage of the machine learning lifecycle. People define the tasks that machine learning systems are intended to address and decide whether or not to build them. The datasets with which machine learning models are trained are generated by people—perhaps explicitly if labels are crowdsourced, or implicitly if they contain traces of people's words, images, or actions. People make decisions about how to collect, clean, and annotate data points, which machine learning models to use (say, a decision tree or a neural network), which training algorithms to implement, and how to incorporate trained models into larger systems. Once developed, people often use the predictions of machine learning systems to make decisions, and these decisions may, in turn, impact other people's lives—potentially in high-stakes domains like criminal justice and healthcare.

Because of the central role that people play in the machine learning lifecycle, building machine learning systems that are reliable, trustworthy, and fair requires that relevant stakeholders—including developers, users, and ultimately the people who are affected by these systems—have at least a basic understanding of how they work. How does a ma-

This is a draft chapter to appear in the upcoming volume tentatively titled *Machines We Trust: Getting Along with Artificial Intelligence* edited by Marcello Pelillo and Teresa Scantamburlo.

chine learning model use different features? Why does a system make certain predictions? How and from where was the dataset with which the model was trained collected? How do these factors influence how well a system's predictions will generalize to new settings, and therefore where the system should be deployed? Being able to answer these questions is more urgent than ever with the widespread movement to "democratize" machine learning (Hosanagar and Saxena, 2017). There is a push to develop off-the-shelf models and tools that make it possible for anyone to incorporate machine learning into their own systems, with or without any prior machine learning experience. However, this leads to the risk that the people who are building, deploying, and affected by machine learning systems may not be used to reasoning about the forms of uncertainty inherent in their predictions, and may therefore misunderstand, mistrust, or misuse these systems. They may fail to account for a system's limitations, be blind to its biases, or be unable to diagnose or debug system failures. And, even if a system behaves as intended, it may still lead to unintended consequences. As Cabitza argues in Chapter X, people's interactions with machine learning systems are "more relevant than computation for [their] impact in real-world settings."

Transparency and intelligibility are often touted as key factors in building trustworthy machine learning systems, yet there is no clear consensus on what these terms mean. Indeed, they are often used to cover a collection of related but distinct concepts. Following recommendations put forth by the European Commission's High-Level Experts Group on Artificial Intelligence, we break transparency into three components (HLEG, 2019):

- **Traceability:** Those who develop or deploy machine learning systems should clearly document their goals, definitions, design choices, and assumptions.
- **Communication:** Those who develop or deploy machine learning systems should be open about the ways they use machine learning technology and about its limitations.
- **Intelligibility:**[1] Stakeholders of machine learning systems should be able to understand and monitor the behavior of those systems to the extent necessary to achieve their goals.

The focus of this chapter is intelligibility. Machine learning researchers have proposed myriad new techniques for achieving intelligibility, many of which we outline in Section 1.3. However, despite this flurry of activity, there is no clear consensus on what makes a system intelligible. Lipton (2018) laid out and contrasted some of the criteria that one might consider, such as simulatability. Doshi-Velez and Kim (2017) pointed out the difficulty of evaluating how intelligible a system is, and the resulting propensity for a "you'll

[1] The High-Level Experts Group refers to this component as *explainability*, while other have referred to it as *interpretability*. Researchers, practitioners, and policy makers have yet to reach consensus on the definitions of these terms, and they are often, though not always, used interchangeably. Throughout this chapter, we stick with *intelligibility*, which dates back to Bellotti and Edwards (2001) and is more often used in the human–computer interaction community (e.g., Lim et al., 2009; Lim and Dey, 2011a,b; Knowles, 2017; Abdul et al., 2018), because it emphasizes the importance of providing particular stakeholders with an understanding of machine learning systems.

know it when you see it" attitude. Citing decades of research in philosophy, psychology, and cognitive science, Miller (2019) argued that the the machine learning community should move away from imprecise, subjective notions of intelligibility, such as whether a system's developer found it "easy to use," and instead evaluate explanations of machine learning systems' behavior in terms of factors known to make explanations more useful to people, such as preferring explanations that are contrastive (e.g., fever is more consistent with a diagnosis of pneumonia as opposed to a common cold) and sequential (e.g., fever suggests either bronchitis or pneumonia and, between the two, chills suggest the latter).

In this chapter, we take the perspective that in order to achieve intelligibility, we must go one step further. Importing existing findings from the social science literature is necessary but not sufficient to ensure that machine learning systems are intelligible. Instead, we must actively engage with the social sciences and—drawing on methodological tools from human–computer interaction (HCI) like interviews and field studies—take a human-centered strategy. We must start by considering the needs of relevant stakeholders, and then design intelligibility techniques with these needs in mind. We must be aware of stakeholders' mental models of machine learning systems. We must evaluate intelligibility techniques in terms of whether these stakeholders are able to achieve their desired goals. Finally, to claim intelligibility, we must posit concrete, testable hypotheses about these stakeholders and their goals and then test these hypotheses experimentally. This strategy makes clear why imprecise, subjective notions of intelligibility are insufficient. Indeed, we discuss scenarios in which commonly accepted assumptions about intelligibility are wrong.

We also argue that model intelligibility is just one piece of a bigger picture. Although model intelligibility has received the most attention in the machine learning literature to date, intelligibility may also be needed for other components of machine learning systems. Depending on the particular stakeholders identified and their desired goals, the intelligibility of components such as datasets and performance metrics may be more important than the intelligibility of models. We therefore give several examples of recent attempts to increase and evaluate the intelligibility of machine learning components other than models.

The intelligibility of machine learning systems is still a relatively young area of research, and in many ways this chapter provides more questions than answers. That said, we hope it will serve as a valuable guide to the intelligible machine learning landscape, as well as a call to action to encourage the machine learning community to come together with researchers from human-centered fields like HCI to form meaningful collaborations.

1.2 Who Needs Intelligibility and Why?

As we argued above, a human-centered strategy to intelligibility must begin with the needs of relevant stakeholders, including data scientists, developers, designers, program managers, regulators, users, or people who are affected by a system, to name just a few examples. Intelligibility is often needed in order to achieve other objectives, such as debugging

a model, anticipating how a system will behave when deployed in the real world, or evaluating whether a system's predictions will lead to decisions that are fair. In this section, we give several illustrative examples of the ways in which needs for intelligibility can arise.

Within the machine learning community, the need for intelligibility perhaps arises most commonly from the goal of improving the robustness of machine learning systems by making it easier for data scientists and developers to identify and fix bugs. For example, Caruana et al. (2015) told the story of a project in the 1990s that was designed to evaluate the application of machine learning to healthcare tasks. In particular, researchers wanted to predict the risk of death for patients with pneumonia so that high-risk patients could be hospitalized. They tried a wide variety of different machine learning models, including a simple rule-based decision support model. Because the rules were human-readable, the researchers were able to examine them directly. They noticed that the model had an odd characteristic: it predicted that having asthma *decreased* a patient's risk of death. At first they were perplexed by this, but eventually they realized that it was an artifact of the dataset with which the model had been trained. Specifically, asthma is known to be a major risk factor, so people with asthma who get pneumonia are immediately admitted to the ICU. As a result, they receive excellent care and tend to have good outcomes. These good outcomes were reflected in the dataset. Because the researchers had used a sufficiently simple model, they were able to identify and fix this bug, which might not have been caught otherwise.

Data scientists and developers may also use intelligibility as a way to gain buy-in from customers or management. In interviews with practitioners in the public sector, Veale et al. (2018) found that one analytics lead working at a tax agency felt pressure to "provide the logic of their machine learning systems" to customers, while another mentioned similar pressure to explain predictions to business users. In both cases, the practitioners therefore decided to forego more complex machine learning models in favor of simpler models.

Outside of the machine learning community, the need for intelligibility typically arises from goals that are more oriented toward users or people who are affected by machine learning systems. For example, O'Neil (2016) described a scenario in which a teacher evaluation system predicted that a teacher's quality had changed dramatically from one year to the next. Providing school administrators and the teacher in question with explanations for these predictions, including descriptions of the most important features, would have allowed these stakeholders to better understand how the system worked, in turn helping them decide how much to trust it. In this case, the system relied only on students' test scores. Had school administrators been provided with explanations that made this clear, they could have decreased their trust in the system accordingly. In other words, such explanations could have served as evidence against the system's trustworthiness (Knowles, 2017).

Intelligibility can also lead to more usable products. For example, providing a user with an explanation of why a recommender system suggested a particular movie may help them decide whether to act on the suggestion (McSherry, 2005). Indeed, Tintarev and Masthoff

(2010) laid out seven distinct goals that one might have when providing explanations of recommendations for users, including effectiveness (i.e., helping users make good decisions), efficiency (i.e., helping users make faster decisions), and satisfaction (i.e., increasing users' ease of use or enjoyment). Similarly, Facebook introduced a "why am I seeing this post?" feature to help users better understand the content on their feeds (Sethuraman, 2019).

Another motivation for intelligibility is the need to demonstrate compliance with regulatory obligations. For example, in 2018, the European Commission expanded plans to regulate online platforms to include search engines.[2] Under the proposed regulations, search engine providers would be required to explain how their ranking systems work—i.e., how different sites, which may include the providers' own sites, are ranked—in order to demonstrate that they are not favoring their own interests over those of third parties. As another example, credit decisions must comply with the Fair Credit Reporting Act and the Equal Credit Opportunity Act, both of which require lenders to explain credit decisions. As a result, the financial services industry primarily relies on machine learning models that are simple enough to constitute explanations of their own behavior. In a similar vein, the European Union General Data Protection Regulation[3] requires that people who are affected by automated decision-making systems must be provided with "meaningful information about the logic involved." Wachter et al. (2018) argued that the most effective way to fulfill this requirement is to provide people with counterfactual explanations that specifically describe how they would need to change their data to experience a different decision.

Finally, intelligibility is often motivated as a way for decision makers to uncover fairness issues in machine learning systems. Although compelling as a motivation, this area has seen less attention to date than one might expect. Some exceptions include the work of Alvarez-Melis and Jaakkola (2017), who proposed a framework for explaining structured predictions and then used this framework to uncover gender stereotypes in machine translation systems, and the work of Tan et al. (2018), who developed a method for using simpler models to audit more complex risk-scoring models in finance and criminal justice for unfairness. Dodge et al. (2019) also showed (among other things) that different types of fairness issues may be more effectively uncovered via different types of explanations. Because of the importance of fairness in machine learning, we recommend the relationship between intelligibility and fairness as a particularly important avenue for future investigation—all the more so because Kleinberg and Mullainathan (2019) showed that, contrary to assumptions, model simplicity and fairness can be at odds with one another.

[2] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51803

[3] https://gdpr-info.eu/

1.3 Common Approaches to Intelligibility

In recent years, machine learning researchers have proposed many new techniques for achieving intelligibility. The majority of these techniques take one of two common approaches. The first common approach is to design and deploy models that are arguably simple enough for people to plausibly reason about their behavior, either directly or with the aid of simple visualizations or other tools—i.e., the model itself constitutes an explanation of its behavior. For example, Ustun and Rudin (2016) and Jung et al. (2017) explored the use of point systems that are simple enough for a decision maker to memorize and apply on the fly, perhaps even without the need for pencil and paper. As another example, GAMs (generalized additive models) can represent relatively more complex functions, but with an additive structure that allows people to visualize the impact of each individual feature on the model's predictions (Lou et al., 2012, 2013; Caruana et al., 2015). Despite the commonly accepted assumption that relatively simple models are less accurate than more complex models like neural networks, there is significant evidence that there are many tasks for which simple models are nearly as accurate as more complex models—including some tasks in high-stakes domains like criminal justice and healthcare—making them a viable alternative when intelligibility is a priority (Dawes, 1979; Astebro and Elhedhli, 2006; Caruana et al., 2015; Jung et al., 2017; Rudin and Ustun, 2018; Rudin, 2019). Therefore, for these tasks, there may be no need to sacrifice accuracy in order to achieve intelligibility.

The second common approach to intelligibility is to generate post-hoc explanations for complex models. Some techniques directly estimate a notion of the "importance" of each feature in generating a prediction for a particular data point. For example, LIME (local interpretable model-agnostic explanations) does this by learning a relatively simple (linear) local approximation of the underlying model around a particular data point (Ribeiro et al., 2016). In contrast, SHAP (Shapley additive explanations) uses a notion of importance based on the idea of Shapley values from cooperative game theory to assign "credit" across all features for a prediction (Lundberg and Lee, 2017). TCAV (testing with concept activation vectors) is a technique that outputs a prediction's sensitivity to a higher-level concept (e.g., whether an object in an image is "striped") that is learned from user-provided examples (Kim et al., 2018). Other techniques aim to explain each prediction in terms of the most influential training data points (Koh and Liang, 2017). Still others provide counterfactual explanations that describe how a data point would need to change to receive a different prediction (Russell, 2019; Weld and Bansal, 2019; Ustun et al., 2019; Wachter et al., 2018). In our own work with collaborators (Alvarez-Melis et al., 2019), we have begun to explore techniques that emphasize factors known to make explanations more useful to people.

With a few exceptions—namely techniques that involve training a simple model to mimic a more complex one (Lakkaraju et al., 2019; Tan et al., 2018)—these post-hoc techniques tend to provide *local* explanations of individual predictions, as opposed to *global* explanations that describe the behavior of a model as a whole. These explanations may or may

not reflect the true behavior of the model (Rudin, 2019; Weld and Bansal, 2019). In a sense, these techniques can be viewed as providing post-hoc justifications for predictions rather than causal explanations of why a model made the predictions it did. Indeed, researchers have shown that some techniques for generating explanations can be prone to instability, failing to output similar explanations for similar predictions made for similar data points (Alvarez-Melis and Jaakkola, 2018a). For these reasons, researchers have begun to explore a hybrid approach of training complex models that are constrained to make only easily explainable predictions (Lei et al., 2016; Alvarez-Melis and Jaakkola, 2018b).

Although explanations can be used to achieve intelligibility, they do not guarantee that intelligibility will be achieved. Moreover, because machine learning systems have many different stakeholders, each of whom may have different goals in different settings, it is seldom the case that there is a single, universal answer as to what makes a system or explanation intelligible. Some techniques may be especially well suited to certain stakeholders. For example, if a system recommends that an individual be denied a loan, the individual may prefer a counterfactual explanation that offers actionable recourse (Ustun et al., 2019). However, this type of explanation may be less useful for a data scientist attempting to debug a model. For this reason, we argue that the design or use of any intelligibility technique must start with an investigation of which stakeholders need intelligibility and why.

1.4 The Importance of Mental Models

Identifying the right intelligibility technique for a particular stakeholder is far from straightforward. Prior work shows that explanations of a system's behavior can be perceived as a waste of time (Bunt et al., 2012), can lead to over- or under-reliance on the system (Bussone et al., 2015), and can even be harmful when the system is uncertain (Lim and Dey, 2011b). In our own research with collaborators (Poursabzi-Sangdeh et al., 2018), we found that commonly accepted assumptions about the relationship between simplicity and intelligibility do not always hold. That is, our intuition about intelligibility can be wrong.

We set out to measure how varying properties of a model that are commonly associated with intelligibility—specifically, the number of features used by the model and the transparency of the model's internals—would influence the behavior of participants who were asked to use the model to make predictions. The participants—laypeople recruited from the crowdsourcing platform Amazon Mechanical Turk—were asked to predict the prices of apartments with the help of a model. For each apartment, the participants were first asked to guess what the model would predict. They were then given the model's true prediction and asked how much they thought the apartment would sell for. We considered four experimental conditions in a 2×2 design. Specifically, we varied whether the model used two features or eight, and whether the model was displayed as a "black box" or as a transparent linear regression model with weights visible to the participants. Crucially, in all four experimental conditions, the participants saw the same model input (i.e., the same apartment features)

and the same model output (i.e., the same predictions). The only differences were in the presentation of the model itself, meaning that any differences in the behavior of the participants between the conditions could be safely attributed to these presentation differences.

Not surprisingly, participants who were shown the transparent, two-feature model were best able to simulate (i.e., guess) the model's predictions. However, contrary to what one might expect, we found no significant difference across the conditions in the degree to which the participants followed the model's predictions for typical data points, when it would have been benecial for them to do so. Even more surprisingly, increased transparency reduced the participants' ability to detect when the model had made a sizable mistake and correct for it. Evidence suggests this was likely due to information overload.

In general, despite the considerable amount of attention given to intelligibility techniques in the machine learning literature, there has been little evaluation of whether these techniques actually help stakeholders achieve their intended goals. This is perhaps unsurprising. User studies of intelligibility techniques are extremely challenging. They require expertise in both machine learning and HCI; they benefit from both qualitative analyses to understand the nuances of how techniques are used in context and quantitative methods to yield scalable findings; they must separate out effects of the model, the intelligibility technique, and the user interface of the specific intelligibility tool; and they must mimic realistic settings without overburdening participants. This last requirement is particularly difficult to satisfy when the stakeholders of interest are specialists, such as doctors using a machine learning system to make healthcare decisions or data scientists trying to debug a model.

In recent work with collaborators (Kaur et al., 2019), we investigated the extent to which two popular intelligibility techniques, GAMs and SHAP (described above in Section 1.3), are effective at facilitating understanding of machine learning models for one key stakeholder group: data scientists. To do this, we conducted a contextual inquiry—semi-structured interviews in which the participants are placed in context—followed by a survey of data scientists to observe how they used tools implementing GAMs and SHAP to uncover common issues that arise when building a model. We found that the participants often over-trusted and misused the tools. Moreover, despite their enthusiasm for the tools, few of the participants were able to accurately describe the tools' visualizations in detail. These results suggest that data scientists may not use intelligibility tools in intended ways.

This gap between what the designers of intelligibility tools intended and the ways in which these tools are perceived and used stems from the fact that users of any computer system are not just passive consumers of information, but rather active partners, who will form their own mental models of the system as they make sense of it (Norman, 1987; Johnson-Laird, 1983; Gentner and Stevens, 1983). One of the first papers to explicitly address the role of mental models in people's interactions with machine learning systems is that of Stumpf et al. (2009). The authors conducted a think-aloud study to assess whether different explanations of a machine learning system would enable users to form useful

mental models, thereby influencing the type of feedback given by the users to improve system performance. Although there has been some recent work in this space (e.g., Kulesza et al., 2013), comparatively more research has focused on using iterative design methods to build better intelligibility tools without directly engaging with mental models. Moreover, people's mental models and system use are influenced by their social context, which is shared via communication. Therefore, if we want people to develop better mental models of machine learning systems, we likely need to design intelligibility techniques that facilitate back-and-forth communication between people and systems (Barnlund, 2008).

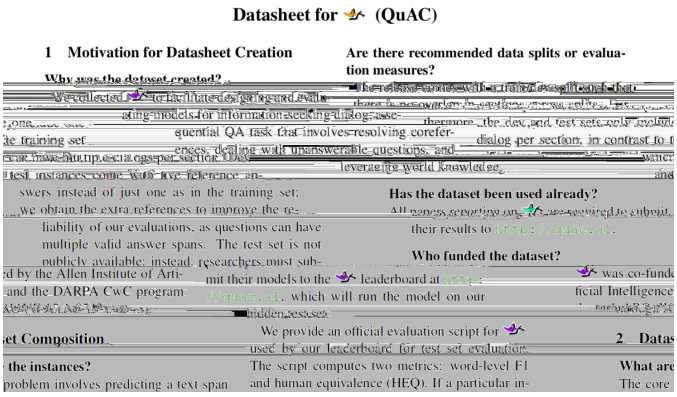
One way to facilitate such back-and-forth communication might be via interactive (rather than static) explanations. For example, Lim et al. (2009) studied techniques for interactively probing a machine learning system to improve intelligibility, while Vaccaro et al. (2018) showed that giving people an interactive slider made them feel more satisfied with the output of a system, supporting the idea that interactivity can give people a sense of agency when interacting with machine learning systems. Weld and Bansal (2019) envisioned interactive explanation systems that allow the user to drill down on specific explanations that are presented, perhaps choosing to view and compare the output from a selection of different intelligibility techniques. Such interactive explanation systems could potentially help with explorative experimentation, as proposed in Chapter X. We highlight interactive intelligibility techniques as an important direction for future investigation.

Finally, we note that in exploring stakeholders' mental models, it may be useful to make the distinction between structural mental models, which help people understand how a system works, and functional mental models, which help people use a system without necessarily understanding how it works (Kulesza et al., 2012). For example, our finding that data scientists were seldom able to accurately describe intelligibility tools' visualizations at more than a superficial level (Kaur et al., 2019) suggests that data scientists may have (at best) functional mental models of these tools. Which stakeholders require structural versus functional mental models of intelligibility tools in which settings remains an open question.

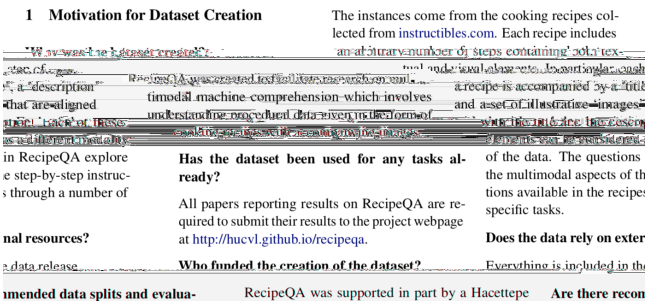
1.5 Beyond Model Intelligibility

Despite the growing body of literature on intelligible machine learning, most papers focus only on the intelligibility of (trained) models. However, a model is just one component of a machine learning system. Depending on the relevant stakeholders and their needs, the intelligibility of other components—such as datasets, training algorithms, or performance metrics—may be even more important than model intelligibility. For example, intelligibility around the composition of training datasets may lead to more actionable insights for a developer attempting to mitigate fairness issues in a system. In this section, we highlight several nascent threads of research on intelligibility beyond machine learning models.

Datasets play a key role in the training and evaluation of machine learning systems. Understanding the relevant aspects of a dataset's origin and characteristics can help people



(a)



(b)

Figure 1.1
Snippets of datasheets created by (a) Choi et al. (2018) and (b) Yagcioglu et al. (2018).

better understand the behavior of any models trained with that dataset. To boost the intelligibility of datasets, Google released Facets, an open-source visualization tool, intended to help people understand and analyze their datasets at different granularities.[4] Taking a different approach, inspired by datasheets for electronic components, we, along with our collaborators, proposed that every dataset be accompanied by a datasheet documenting its motivation, composition, collection, pre-processing, potential uses, distribution, and maintenance (Gebru et al., 2018). Datasheets can help dataset creators uncover hidden

[4] https://pair-code.github.io/facets/

assumptions, and can help dataset consumers (such as model developers reusing existing datasets) determine if a particular dataset meets their needs. Although there is still significant research that needs to be done to understand how effective these approaches are for these different stakeholders, dataset intelligibility has incredible potential to uncover possible sources of unfairness or broader ethical issues. Furthermore, it can reveal problems with a dataset before that dataset is used to train a model—i.e., at a stage of the machine learning lifecycle during which data-related problems can be more easily addressed.

Moving beyond datasets, in 2019, the Partnership on AI launched the ABOUT ML initiative,[5] which aims to develop best practices for increasing the transparency of machine learning components via documentation. This initiative draws on existing proposals for documenting datasets (Gebru et al., 2018; Bender and Friedman, 2018), models (Mitchell et al., 2019), and entire machine learning systems (Arnold et al., 2018). Efforts like ABOUT ML have the potential to increase the intelligibility of every machine learning component for a wide variety of stakeholders, though again, more research is needed to understand which practices will best help which stakeholders achieve their desired goals. For example, a customer deciding whether to purchase a machine learning system may benefit less from detailed datasheets for the datasets with which the system was trained than a higher-level document outlining the system's capabilities and limitations, similar to the "transparency note" recently published by Microsoft about its facial recognition API.[6]

As another example, users of machine learning systems, including doctors, judges, and other decision makers, need to understand performance metrics, such as accuracy, precision, or recall on held-out data, in order to decide how much to trust a model's predictions. Together with our collaborator, we conducted large-scale human-subject experiments to determine whether laypeople's trust in a machine learning model varied depending on the model's stated accuracy on a held-out dataset and on its observed accuracy in practice (Yin et al., 2019). Operationalizing trust in multiple ways, including self-reported levels of trust and the frequency with which the participants adjusted their predictions to match those of the model, we found that stated accuracy and observed accuracy both affected the participants' trust in the model. Moreover, the effect of stated accuracy varied depending on the model's observed accuracy. In particular, participants were more likely to increase their trust in the model if the model's observed accuracy was higher than their own accuracy.

In a similar vein, Kocielnik et al. (2019) studied the effect of laypeople's expectations about a machine learning system on their subjective perceptions of accuracy and their acceptance of the system. They found that different techniques for setting expectations, including stating the system's accuracy, were effective at making expectations more re-

[5] https://www.partnershiponai.org/about-ml/

[6] https://azure.microsoft.com/en-us/resources/transparency-note-azure-cognitive-services-face-api/

alistic and increasing participants' acceptance of an imperfect machine learning system, depending on the types of errors made by the system. Together, these results highlight the importance of responsibly communicating information about performance metrics to users, because this information can influence the ways in which they interact with a system.

Finally, some stakeholders need intelligibility of the errors made by a system, in addition to aggregate performance metrics. For example, developers need to understand a model's errors in order to debug it, while customers need to understand a model's errors in order to decide whether a particular deployment context is appropriate or not. To that end, several tools for developing fairer machine learning systems—including Aequitas,[7] AI Fairness 360,[8] and fairlearn,[9] among others—provide ways to disaggregate performance metrics by known pivots such as demographic groups. Similarly, Barraza et al. (2019) developed an error analysis tool to help people explore the terrain of errors made by a machine learning system. The tool provides users with different views of the system's errors, enabling them to examine these errors grouped by known pivots or grouped by learned clusters. As a case study, the authors used the tool to explore the errors made by a commercial gender classifier. They found that it had particularly high error rates for images of women with short hair who were not wearing make-up and were not smiling. As another example, both Amershi et al. (2015) and Ren et al. (2016) designed visualization tools that enable easier error examination and debugging compared to using aggregate performance metrics. Error intelligibility can help developers make smarter choices about which data to collect, features to include, or objective functions to optimize in order to improve system performance.

1.6 What Machine Learning Researchers Can Learn From HCI

Taking a human-centered strategy to intelligible machine learning raises questions about when and how human-centered fields—especially HCI, a long-established field focusing on the ways in which people interact with computers—should play a role (cf. Abdul et al., 2018). We argue that intelligible machine learning should not be a viewed as disciplinary subfield of machine learning, as it is currently sometimes positioned, but should instead be viewed as a genuinely interdisciplinary area that demands tight integration between the machine learning and HCI communities. Indeed, the term "intelligibility" has a long history in HCI, where Bellotti and Edwards (2001) defined it in terms of systems that "represent to their users what they know, how they know it, and what they are doing about it." There is already some work that advocates for drawing on the HCI literature when designing or studying intelligibility techniques (e.g., Miller, 2019), but tight integration demands that we go further, prioritizing collaborations between machine learning and HCI researchers.

[7] `http://www.datasciencepublicpolicy.org/projects/aequitas/`

[8] `https://aif360.mybluemix.net/`

[9] `https://github.com/fairlearn/fairlearn`

Standard methodological tools typically employed in HCI, which include human-subject experiments, surveys, and ethnographic inquiries, are critical for moving away from imprecise, subjective notions of intelligibility. For example, as we illustrated above, human-subject experiments can reveal that commonly accepted assumptions about intelligibility are wrong. Use of these methodological tools, many of which are drawn from and shared with other human-centered fields such as psychology and anthropology, will enable researchers to develop a more comprehensive and well-founded understanding of how to help different stakeholders achieve their desired goals. Indeed, beyond our own work (described above in Section 1.4), others in the machine learning community have begun to use human-subject experiments to evaluate intelligibility techniques (e.g., Lage et al., 2018, 2019).

As one example of how this can be done, Hohman et al. (2019) used an iterative co-design process to develop an interactive visualization tool on top of GAMs to help data scientists better understand their models. The authors started with a need-finding study, and then used the findings from this study to develop and evaluate the visualization tool. Although the evaluation involved subjective assessments of the tool, these assessments were made by the participants in the study—in this case, data scientists—rather than the authors.

There is also much that the machine learning community can learn from the wide range of theories about human cognition and social interaction that have been proposed in the HCI and social science literature. For example, in Section 1.4, we argued that interactive explanations may potentially play an important role in helping stakeholders develop better mental models of machine learning systems. As pointed out by Kaur et al. (2019), to explore the benefits of interactivity, it may be worth viewing intelligibility techniques through the lens of communication theory. Transactional models of communication (Barnlund, 2008) that incorporate verbal, non-verbal, and behavioral cues might suggest new ways of taking a human-centered strategy when designing or studying intelligibility techniques.

We see tight integration between the machine learning and HCI communities as necessary to the success of intelligible machine learning. Active engagement of this sort will make sure that research on intelligibility is situated in a broader human-centered context, that intelligibility techniques are not based on intuition alone, and that any findings that are relevant to HCI beyond the context of intelligibility make their way into the HCI literature.

1.7 Summary

Prompted by the widespread use of machine learning systems throughout society, there is an increasing demand to make these systems intelligible. In response, machine learning researchers have proposed a wide variety of techniques for achieving intelligibility, including models that are simple enough to constitute explanations of their own behavior and post-hoc explanations for potentially complex models. However, many of these techniques are based on commonly accepted assumptions about intelligibility, which can be wrong. In this chapter, we argued that intelligibility is a fundamentally human-centered concept and must

be treated as such. The needs of relevant stakeholders must be actively considered from start to finish when designing intelligibility techniques. To do this, machine learning researchers must actively engage with the social sciences—drawing on methodological tools from HCI and other human-centered fields, and fostering interdisciplinary collaborations.

We have only just scratched the surface of intelligible machine learning. Many directions remain to be explored. We highlight two as being especially important: 1) better understanding who the relevant stakeholders are, what their desired goals are, and how intelligibility techniques can be best designed to help them achieve these goals, and 2) moving beyond the model to bring intelligibility to other components of machine learning systems.

Acknowledgments

We thank David Alvarez-Melis, Saleema Amershi, Jacquelyn Krones, Edith Law, Scott Lundberg, Vidya Muthukumar, and Ming Yin for providing feedback on earlier versions of this chapter. This chapter also greatly benefited from our conversations with Solon Barocas, Rich Caruana, Hal Daumé, Dan Goldstein, Jake Hofman, Harman Kaur, Sam Jenkins, Harsha Nori, Mike Philips, Forough Poursabzi-Sangdeh, and Abi Sellen, as well as all past and present members of Microsoft's Aether Working Group on Intelligibility and Explanation. Thanks to Marcello Pelillo and Teresa Scantamburlo for making this book happen.

Bibliography

Abdul, Ashraf, 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 (CHI)*.

Alvarez-Melis, David, and Tommi S. Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP)*, 412–421.

Alvarez-Melis, David, and Tommi S. Jaakkola. 2018a. On the robustness of interpretability methods. In *ICML workshop on human interpretability in machine learning*.

Alvarez-Melis, David, and Tommi S. Jaakkola. 2018b. Towards robust interpretability with self-explaining neural networks. In *Advances in neural information processing systems 31 (NeurIPS)*, 7775–7784.

Alvarez-Melis, David, Hal Daumé, III, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Weight of evidence as a basis for human-oriented explanations. In *NeurIPS workshop on human-centered machine learning*.

Amershi, Saleema, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 2015 CHI conference on human factors in computing systems (CHI)*, 337–346.

Arnold, Matthew, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2018. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. CoRR arXiv:1808.07261.

Astebro, Thomas, and Samir Elhedhli. 2006. The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science* 52 (3): 395–409.

Barnlund, Dean C. 2008. A transactional model of communication. In *Communication theory*, ed. C. David Mortensen, 47–57. Taylor & Francis.

Barraza, Rick, Russell Eames, Yan Esteve Balducci, Josh Hinds, Scott Hoogerwerf, Eric Horvitz, Ece Kamar, Jacquelyn Krones, Josh Lovejoy, Parham Mohadjer, Ben Noah, and Besmira Nushi. 2019. Error terrain analysis for machine learning: Tool and visualizations. In *ICLR workshop on debugging machine learning models*.

Bellotti, Victoria, and Keith Edwards. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction* 16 (2–4): 193–212.

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

Bunt, Andrea, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on intelligent user interfaces (IUI)*, 169–178.

Bussone, Adrian, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the IEEE international conference on healthcare informatics (ICHI)*.

Caruana, Rich, 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 21st ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, 1721–1730.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. CoRR arXiv:1808.07036.

Dawes, Robyn M. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* 34 (7): 571–582.

Dodge, Jonathan, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces (IUI)*, 275–285.

Doshi-Velez, Finale, and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. CoRR arXiv:1702.08608.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. CoRR arXiv:1803.09010.

Gentner, Dedre, and Albert L. Stevens. 1983. *Mental models*. Lawrence Erlbaum Associates.

High-Level Expert Group on Artificial Intelligence (HLEG). 2019. Ethics guidelines for trustworthy AI. Retrieved from `https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines`.

Hohman, Fred, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI)*.

Hosanagar, Kartik, and Apoorv Saxena. 2017. The democratization of machine learning: What it means for tech innovation. Retrieved from `http://knowledge.wharton.upenn.edu/article/democratization-ai-means-tech-innovation/`.

Johnson-Laird, Philip. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge University Press.

Jung, Jongbin, Connor Concannon, Ravi Shro, Sharad Goel, and Daniel G. Goldstein. 2017. Simple Rules for Complex Decisions. CoRR arXiv:1702.04690.

Kaur, Harmanpreet, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2019. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. Working paper.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th international conference on machine learning (ICML)*, 2668–2677.

Kleinberg, Jon, and Sendhil Mullainathan. 2019. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM conference on economics and computation (EC)*, 807–808.

Knowles, Bran. 2017. Intelligibility in the face of uncertainty. In *CHI workshop on designing for uncertainty in HCI*.

Kocielnik, Rafal, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI)*.

Koh, Pang Wei, and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th international conference on machine learning (ICML)*, 1885–1894.

Kulesza, Todd, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI)*, 1–10.

Kulesza, Todd, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Proceedings of the IEEE symposium on visual languages and human-centric computing*, 3–10.

Lage, Isaac, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2018. Human-in-the-loop interpretability prior. In *Advances in neural information processing systems 31 (NeurIPS)*, 10159–10168.

Lage, Isaac, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the 7th AAAI conference on human computation and crowdsourcing (HCOMP)*, 59–67.

Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM conference on artificial intelligence, ethics, and society (AIES)*, 131–138.

Lei, Tao, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 conference on empirical methods in natural language processing (EMNLP)*, 107–117.

Lim, Brian Y., and Anind K. Dey. 2011a. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services (MobileHCI)*, 157–166.

Lim, Brian Y., and Anind K. Dey. 2011b. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on ubiquitous computing (UbiComp)*, 415–424.

Lim, Brian Y., Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI)*, 2119–2128.

Lipton, Zachary C. 2018. The mythos of model interpretability. *Communications of the ACM* 61 (10): 36–43.

Lou, Yin, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, 150–158.

Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, 623–631.

Lundberg, Scott, and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30 (NIPS)*, 4765–4774.

McSherry, David. 2005. Explanation in recommender systems. *Artificial Intelligence Review* 24 (2): 179–197.

Miller, Tim. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artifical Intelligence* 267: 1–38.

Mitchell, Margaret, 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 (FAT*)*, 220–229.

Norman, Don A. 1987. Some observations on mental models. In *Human-computer interaction: A multidisciplinary approach*, eds. R. M. Baecker and W. A. S. Buxton, 241–244. Morgan Kaufmann Publishers Inc..

O'Neil, Cathy. 2016. *Weapons of math destruction*. Crown Publishing.

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. CoRR arXiv:1802.07810.

Ren, Donghao, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D. Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 61–70.

Ribeiro, Marco Tulio, 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 (KDD)*, 1135–1144.

Rudin, Cynthia. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (5): 206–215.

Rudin, Cynthia, and Berk Ustun. 2018. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Applied Analytics* 48 (5): 449–466.

Russell, Chris. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the conference on fairness, accountability, and transparency (FAT*)*, 20–28.

Sethuraman, Ramya. 2019. Why am I seeing this? We have an answer for you. Retrieved from `https://newsroom.fb.com/news/2019/03/why-am-i-seeing-this/`.

Stumpf, Simone, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67 (8): 639–662.

Tan, Sarah, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 3rd AAAI/ACM conference on artificial intelligence, ethics, and society (AIES)*.

Tintarev, Nava, and Judith Masthoff. 2010. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook*, eds. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, 479–510. Springer.

Ustun, Berk, and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning Journal* 102 (3): 349–391.

Ustun, Berk, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency (FAT*)*, 10–19.

Vaccaro, Kristen, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI conference on human factors in computing systems (CHI)*.

Veale, Michael, 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 (CHI)*.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31 (2): 841–887.

Weld, Daniel S., and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Communications of the ACM* 62 (6): 70–79.

Yagcioglu, Semih, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. CoRR arXiv:1809.00812.

Yin, Ming, 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 (CHI)*.