CHAPTER 1
Introduction

Explainable AI, also known as *XAI*, is a field of machine learning (ML) consisting of techniques that aim to give a better understanding of model behavior by providing explanations as to how a model made a prediction. Knowing how a model behaves, and how it is influenced by its training dataset, gives anyone who builds or uses ML powerful new abilities to improve models, build confidence in their predictions, and understand when things go awry. Explainable AI techniques are especially useful because they do not rely on a particular model—once you know an Explainable AI method, you can use it in many scenarios. This book is designed to give you the ability to understand how Explainable AI techniques work so you can build an intuition for when to use one approach over another, how to apply these techniques, and how to evaluate these explanations so you understand their benefits and limitations, as well as communicate them to your stakeholders. Explanations can be very powerful and are easily able to convey a new understanding of why a model makes a certain prediction, as Figure 1-1 demonstrates, but they also require skill and nuance to use correctly.

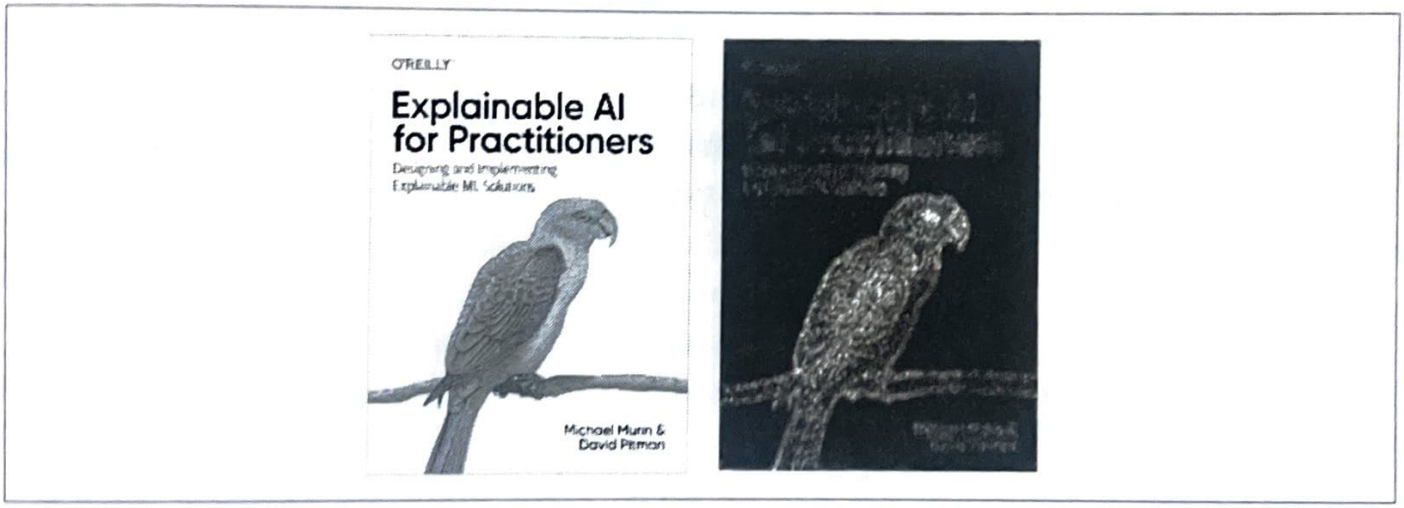

Figure 1-1. An explanation using Blur-IG (described in Chapter 4) that shows what pixels influenced an image classification model to predict that the animal on the cover of this book is a parrot.

Why Explainable AI

In 2018, data scientists built a machine learning (ML) model to identify diseases from chest X-rays. The goal was to allow radiologists to be able to review more X-rays per day with the help of AI. In testing, the model achieved very high, but not perfect, accuracy with a validation dataset. There was just one problem: the model performed terribly in real-world settings. For months, the researchers tried to find why there was a discrepancy. The model had been trained on the same type of chest X-rays shown in Figure 1-2, and the X-rays had any identifying information removed. Even with new data, they kept encountering the same problem: fantastic performance in training, only to be followed by terrible results in a hospital setting. Why was this happening?

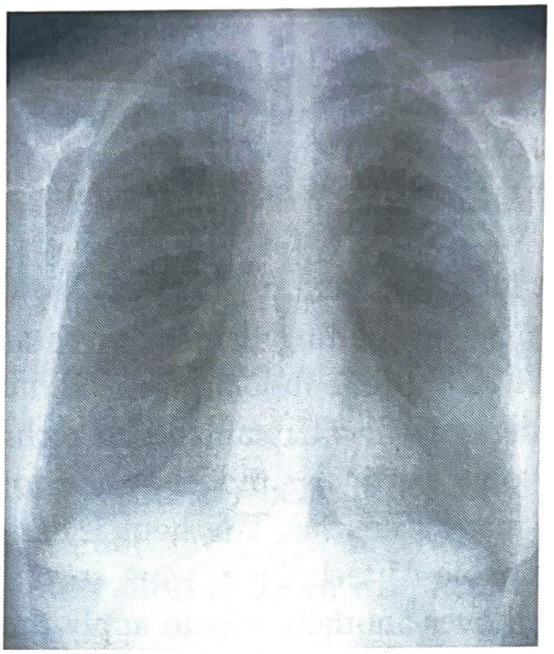

Figure 1-2. An example of the chest X-rays used to train the model to recognize diseases. Can you identify what led the model astray?[1]

A few years later, another research group, this time eye doctors in the UK, embarked on a mission to train a model to identify diseases from retinal scans of a patient's eye (see Figure 1-3). After they had trained the model, they encountered an equally surprising, but very different result. While the model was very good at identifying diseases, it was uncannily accurate at also predicting the sex of the patient.

1 The authors thank the National Cancer Institute (NCI) for access to their data collected by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

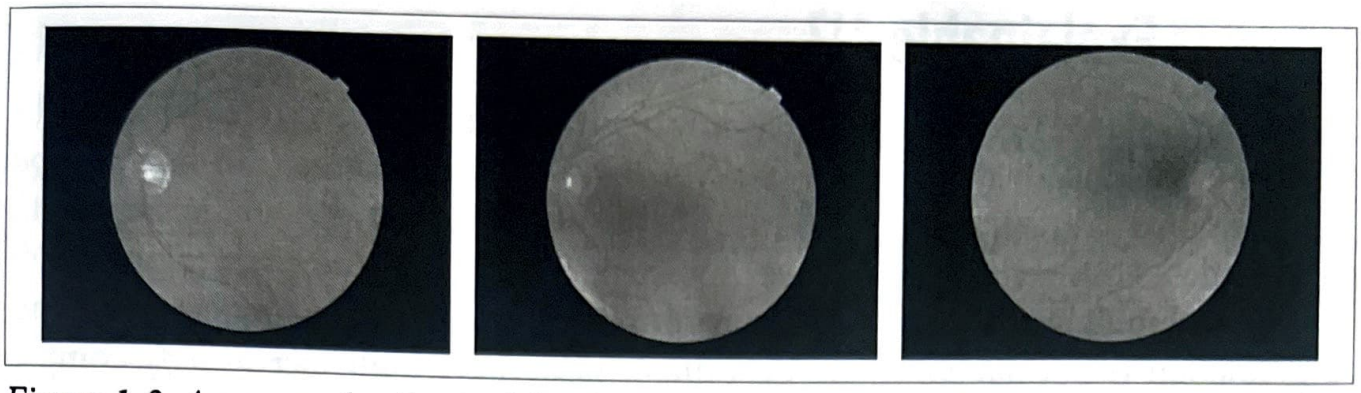

Figure 1-3. An example of retinal fundus images, which show the interior of an eye and can be used to predict diseases such as diabetes.

There were two fascinating aspects of this prediction. First, the doctors had not designed the ML to predict the patient's sex; this was an inadvertent output of their model architecture and ML experiments. Second, if these predictions were accurate, this correlation between the interior of our eyes and our sex was a completely new discovery in ophthalmology. Had the ML made a brand-new discovery, or was something flawed in the model or dataset such that information about a patient's sex was leaking into the model's inference?

In both cases of the chest X-ray and the retinal images, the opaque nature of machine learning had turned on its users. Modern machine learning has succeeded precisely because computers could teach themselves how to perform many tasks using an approach of consuming vast amounts of information to iteratively tune a large number of parameters. However, the large quantities of data involved in training these models have made it pragmatically impossible for a human to directly examine and understand the behavior of a model or how a dataset influenced the model. While any machine learning model can be inspected by looking at individual weights or specific data samples used to train the model, this examination rarely yields useful insights.

In the first example we gave, the X-ray model that performed well in testing and was useless in practice, the data scientists who built the model did all the right things. They removed text labels from the images to prevent the model from learning how to read and predict a disease based on ancillary information. They properly divided their training, testing, and validation datasets, and used a reasonable model architecture that built upon an existing process (radiologists examining X-rays) that already proved it was feasible to identify diseases from a patient's X-ray. And yet, even with these precautions and expertise, their model was still a failure. The eye doctors who built an ML model for classifying eye disease were world-class experts in their own field who understood ophthalmology but were not machine learning experts. If their discovery that our eyes, which have been exhaustively studied for hundreds of years, still held secrets about human biology, how could they perform a rigorous analysis of the machine learning model to be certain their discovery was real? Fortunately, Explainable AI provides new ways for ML practitioners, stakeholders, and end users to answer these types of questions.

What Is Explainable AI?

At Explainable AI's core is the creation of explanations, or to create explainability for a model. Explainability is the concept that a model's actions can be described in terms that are comprehensible to whoever is consuming the predictions of the ML model. This explainability serves a variety of purposes, from improving model quality to building confidence, and even providing a pathway for remediation when a prediction is not one you were expecting. As we have built increasingly complex models, we have discovered that a high-performing model is not sufficient to be acceptable in the real world. It is necessary that a prediction also has a reasonable explanation and, overall, the model behaves in the way its creators intended.

Imagine an AI who is your coworker on a project. Regardless of how well your AI coworker performs any task you give them, it would be incredibly frustrating if your entire collaboration with the AI consisted of them vanishing after taking on a task, suddenly reappearing with the finished work, and then vanishing again as soon as they delivered it to you. If their work was superb, perhaps you would be accepting of this transactional relationship, but the quality of your AI coworker's results can vary considerably. Unfortunately, the AI never answers your questions or even tells you how they arrived at the result.

As AI becomes our coworkers, colleagues, and more responsible for decisions affecting many aspects of our life, the feedback is clear that having AI as a silent partner is unsatisfying. We want (and in the future, will have a right) to expect we can have a two-way dialogue with our machine learning model to understand why it performed the way it did. Explainable AI represents the beginning of this dialogue, by opening up a new way for an ML system to convey how it works instead of simply delivering the results of a task.

Who Needs Explainability?

To understand how explainability aided the researchers in our two examples of where conventional ML workflows failed to address the issues encountered, it is also necessary to talk about who uses explainability and why. In our work on Explainable AI for Google Cloud, we have engaged with many companies, data scientists, ML engineers, and business executives who have sought to understand how a model works and why. From these interactions, we have found that there are three distinct groups of people who are seeking explanations, and each group has distinct but overlapping motivations and needs. Throughout the book, we will refer to all of these groups as *explainability consumers* because they are the recipient of an explanation and act upon that explanation. Our consumers can be divided into three roles:

Practitioners
: Data scientists and engineers who are familiar with machine learning

Observers
: Business stakeholders and regulators who have some familiarity with machine learning but are primarily concerned with the function and overall performance of the ML system

End users
: Domain experts and affected users who may have little-to-no knowledge of ML and are also recipients of the ML system's output

A person can simultaneously assume multiple roles as an ML consumer. For example, a common pattern we see is that data scientists start as ML practitioners, but over time build up an understanding of the field they are serving and eventually become domain experts themselves, allowing them to act as an end user in evaluating a prediction and explanation.

In our chest X-ray case study, the ML practitioners built the model but did not have domain expertise in radiology. They understood how the ML system works and how it was trained but did not have a deep understanding of the practice of radiology. In contrast, in the retina images case study, the ophthalmologist researchers with domain expertise who built the model found the ML had discovered a new correlation between the appearance of the interior of our eyes and our sex, but they lacked the expertise of ML practitioners to be confident the model was functioning correctly.

Each group had very different needs for explaining why the model acted the way it did. The ML practitioners were looking for precise and accurate explanations that could expose the step at which the ML had failed. The ophthalmologists, as end users, were looking for an explanation that was more conceptual and would help them construct a hypothesis for why the classification occurred and also allow them to build trust in the model's predictions.

Challenges in Explainability

How we use Explainable AI for a model depends on the goals of those consuming the explanations. Suppose, based on just the information we have given about these individuals, their ML, and their challenges, we asked you to implement an Explainable AI technique that will generate explanations of how the model arrived at its predictions. You toil away and implement a way to generate what you think are good explanations for these ML models by using Integrated Gradients (Chapter 4) to highlight pixels that were influential in the prediction. You excitedly deliver a set of predictions with relevant pixels highlighted in the image, but rather than being celebrated as the person who saved the day, you immediately get questions like:

- How do I know this explanation is accurate?
- What does it mean that one pixel is highlighted more than another?
- What happens if we remove these pixels?
- Why did it highlight these pixels instead of what we think is important over here?
- Could you do this for the entire dataset?

In trying to answer one question, you inadvertently caused five more questions to be asked! Each of these questions are valid and worth asking, but may not help your audience in their original goal to understand the model's behavior. However, as explainability is a relatively new field in AI, it is likely you will encounter these questions, for which there are no easy answers. In Explainable AI, there are several outstanding challenges:

- Demonstrating the semantic correctness of explanation techniques above and beyond the theoretical soundness of the underlying mathematics
- Combining different explanation techniques in an easy and safe way that enhances understanding rather than generating more confusion
- Building tools that allow consumers to easily explore, probe, and build richer explanations
- Generating explanations that are computationally efficient
- Building a strong framework for determining the robustness of explanation techniques

Promising research is being conducted in all of these areas; however, none have yet achieved acceptance within the explainability community to the level that we would feel confident in recommending them. Additionally, many of these questions have led to research papers that investigate how explanation techniques may be fundamentally broken. This is a very promising line of research but, as we discuss in Chapter 7, may be better viewed as research that probes how susceptible XAI techniques are to adversarial attacks or are brittle and unable to generate good explanations outside of their original design parameters. Many of these questions are sufficiently interesting that we recommend caution when using explanations for high-risk or safety-critical AIs.

Evaluating Explainability

Let's return to our two case studies to see how they fared after using explainability. For the chest X-ray, the ML practitioners had been unable to discover why the model performed very well in training and testing, but poorly in the real world. One of the researchers used an explainability technique, Integrated Gradients, to highlight pixels in the chest X-ray that were influential in the prediction. Integrated Gradients

(covered in more depth in Chapter 4) evaluates the prediction by starting with a baseline image—for example, one that is all black, and progressively generating new images with pixels that are closer to the original input. These intermediate images are then fed to the model to create new predictions. The predictions are consolidated into a new version of the input image where pixels that were influential to the model's original prediction are shown in a new color, which is known as a saliency map. The intensity of the coloring reflects how strong the pixels influenced the model. At first glance, the explanations were as baffling as the original problem.

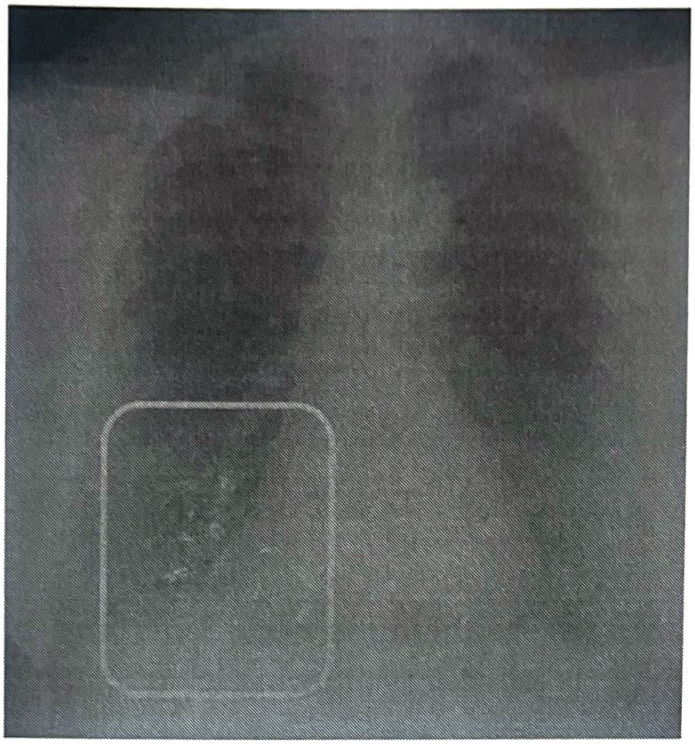

Figure 1-4. The explanation for which pixels, highlighted in red, the model thought indicated a disease in the chest X-ray. For readability, the area of the image containing the most attributed pixels is outlined by the blue box. (Print readers can see the color image at https://oreil.ly/xai-fig-1-4.)

An example of one of these images is shown in Figure 1-4, and it may appear that no pixels were used in the prediction. However, if you look at the lower left of this image, you will notice a smattering of red pixels among the black and white of the chest X-ray. These appear to be quite random as well. It's not until one closely looks at this area of the image that you can barely perceive what appear to be scratch marks on the X-ray. However, these are not random scratches, but the pen marks of a radiologist who had drawn where the disease was in the X-ray, as we can see in Figure 1-5. The model then became trained to associate pen markings with a disease being present. However, in the real-world setting, the X-rays had no pen markings on them because the raw X-ray images were fed to the model before being shown to a radiologist who could have marked them up.

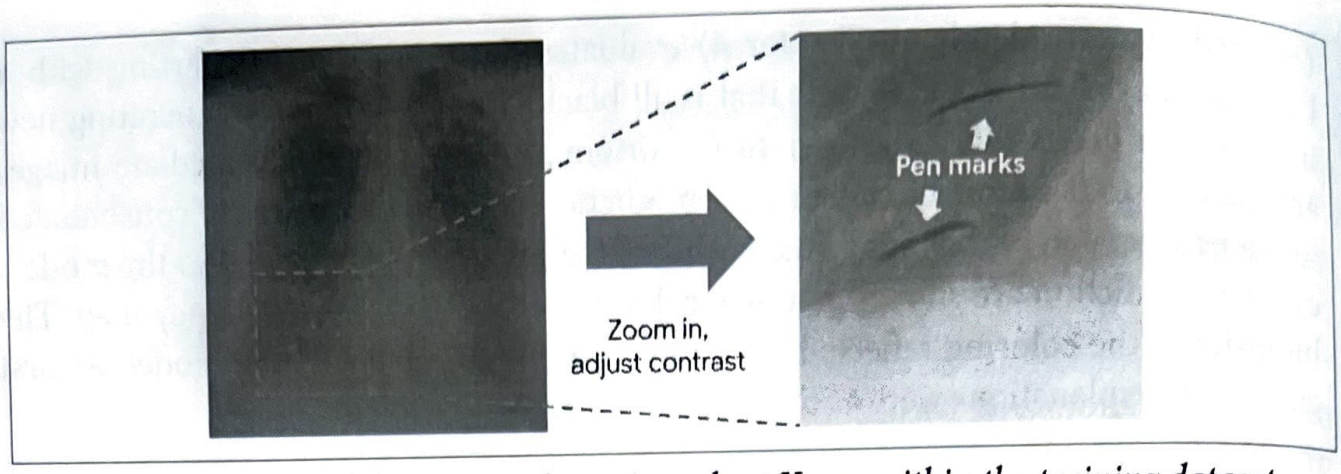

Figure 1-5. Example of the pen markings in a chest X-ray within the training dataset.

Once the researchers figured out the cause of their performance mismatch, that pen markings had leaked information to the model about whether to classify the X-ray as showing disease or not, they could build a new dataset that had not been annotated by radiologists. The model's subsequent performance was not noticeably worse in training than the original model and performed better in the real-world setting.

Let's turn to the retina study. How did ophthalmologists use Explainable AI to become confident that their ML model could predict the sex of a patient based on their retinal fundus images? The researchers used a technique known as XRAI (discussed in Chapter 4; no relation to X-rays) to highlight regions of the eye image that influenced the model's prediction. The explanations, seen in Figure 1-6, showed that the model was attentive to the optic nerve (the large blob to one side of the retina) and the blood vessels radiating out from the optic nerve.

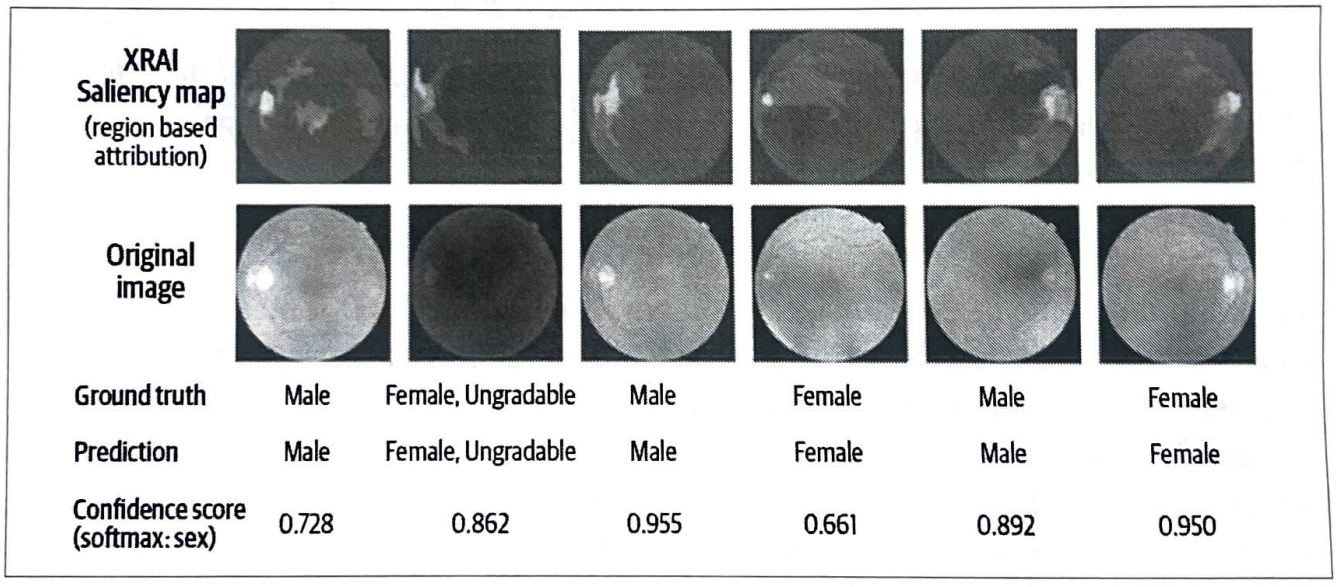

Figure 1-6. XRAI used to highlight what pixels influenced the model's prediction of a patient's sex based on a photograph of the interior of their eye, from an article by Korot et al.[2]

2 Edward Korot et al., "Predicting Sex from Retinal Fundus Photographs Using Automated Deep Learning," *Scientific Reports* 11, article no. 10286 (May 2021), *https://oreil.ly/Le50t*.

By seeing that the model had become influenced by such specific parts of the eye's anatomy, the researchers were convinced that the model was indeed making correct predictions. This work was also sufficient to convince the broader scientific community, as the results were eventually published as a paper in *Scientific Reports*.

How Has Explainability Been Used?

The two examples we gave focused on explainability for image models in medical research and healthcare. You may often find that examples of explainability involve an image model because it is easier to understand the explanation for an image than the relative importance of different features in structured data or the mapping of influential tokens in a language model. In this section, we look at some other case studies of how Explainable AI has been used beyond image models.

How LinkedIn Uses Explainable AI

Since 2018, LinkedIn has successfully used Explainable AI across many areas of its business, from recruiting to sales, and ML engineering. For example, in 2022 LinkedIn revealed that Explainable AI was key to the adoption of a ranking and recommendation ML system used by their sales team to prioritize which customers to engage with based on the ML's prediction of how likely it was that the customer would stop using existing products (also known as *churn*), or their potential to be sold new ones (known as *upselling*). While the ML performed well, the AI team at LinkedIn quickly discovered that the system was not going to be used by their sales teams unless it included a rationale for the predictions:

> While this ML-based approach was very useful, we found from focus group studies that ML-based model scores alone weren't the most helpful tool for our sales representatives. Rather, they wanted to understand the underlying reasons behind the scores—such as why the model score was higher for Customer A but lower for Customer B—and they also wanted to be able to double-check the reasoning with their domain knowledge.[3]

Similar to our chest X-ray example, LinkedIn has also used Explainable AI to improve the quality of their ML models. In this case, their ML team productionized the use of explainability across many models by building a tool that allows LinkedIn data scientists and engineers to perturb features (see Chapter 3) to generate alternative scenarios for predictions to understand how a model may behave with a slightly different set of inputs.[4]

---

3 Jilei Yang et al., "The Journey to Build an Explainable AI-Driven Recommendation System to Help Scale Sales Efficiency Across LinkedIn," LinkedIn, April 6, 2022, *https://oreil.ly/JJPj9*.

4 Daniel Qiu and Yucheng Qian, "Relevance Debugging and Explaining at LinkedIn," LinkedIn, 2019, *https://oreil.ly/cSFhj*.

LinkedIn has gone a step further to create an app, CrystalCandle,[5] that translates raw explanations for structured data, which are often just numbers, into narrative explanations (we discuss narrative explanations further in Chapter 7). An example of the narrative explanations they have built are shown in Table 1-1.

Table 1-1. LinkedIn's CrystalCandle comparison of raw explanations versus the corresponding narrative explanations

Model prediction and interpretation (nonintuitive)	Narrative insights (user-friendly)
Propensity score: 0.85 (top 2%) Top important features (with importance score): • `paid_job_s4`: 0.030 • `job_view_s4`: 0.013 • `hire_cntr_s3`: 0.011 • `conn_cmp_s4`: 0.009	This account is extremely likely to upsell. Its upsell likelihood is larger than 98% of all accounts, which is driven by: • Paid job posts changed from 10 to 15 (+50%) in the last month. • Views per job changed from 200 to 300 (+50%) in the last month.

PwC Uses Explainable AI for Auto Insurance Claims

Working with a large auto insurer, PricewaterhouseCoopers (PwC) built an ML system to estimate the amount of an insurance claim. In building the system, PwC clearly highlights how Explainable AI was not an optional addition to their core project, but a necessary requirement for the ML to be adopted by the insurance company, their claims adjusters, and customers. They call out four different benefits from using explainability in their ML solution:

> The company's explainable AI model was a game changer as it enabled the following:
>
> - empowered auto claim estimators to identify where to focus attention during an assessment
> - provided approaches for sharing knowledge among the estimator team to accurately determine which group should handle specific estimates
> - identified 29% efficiency savings possible with full implementation of proof of concept models across the estimator team
> - reduced rework and improved customer experience through reduced cycle times[6]

In our work with customers at Google Cloud, we have also seen similar benefits to many customers who have built explainability into their AI.

5 Jilei Yang et al., "CrystalCandle: A User-Facing Model Explainer for Narrative Explanations," arXiv, 2021, *https://arxiv.org/abs/2105.12941*.

6 See PwC, "Insurance Claims Estimator Uses AI for Efficiency," (*https://oreil.ly/3vMUr*) for more information.

Accenture Labs Explains Loan Decisions

The experience of receiving a loan from a bank can be a confusing experience. The loan applicant is asked to fill in many forms and provide evidence of their financial situation and history in order to apply for a loan, often only asking for approval for the broad terms of the loan and the amount. In response, consumers are either approved, often with an interest rate decided by the bank, or they are denied with no further information. Accenture Labs demonstrated how even a common loan, the Home Equity Line of Credit (HELOC), could benefit from providing positive counterfactual explanations (covered in Chapter 2) as part of an ML's prediction for whether to approve or deny a loan application. These counterfactual explanations focused on creating a "what-if" scenario for what aspects of the applicant's credit history and financial situation would have resulted in the loan being approved or denied. In this case study, Accenture focused on understanding how explainability could be used across different ML systems,[7] demonstrating the value of how using Explainable AI allowed for explanations to still be generated, while the underlying model was changed to different model architectures.

DARPA Uses Explainable AI to Build "Third-Wave AI"

The Defense Advanced Research Projects Agency (DARPA), an arm of the US Department of Defense, conducted a five-year program (*https://oreil.ly/xAbXI*)[8] with many projects to investigate the use of Explainable AI. DARPA's goals in using Explainable AI are to "produce more explainable models, while maintaining a high level of learning performance" and "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners." DARPA believes explainability is a key component of the next generation of AI systems, where "machines understand the context and environment in which they operate, and over time build underlying explanatory models that allow them to characterize real-world phenomena." Over the past few years, the program has had several annual workshops demonstrating the feasibility of building explainability into many different types of ML, from data analysis to autonomous systems and assistive decision-making tools.

Summary

In this chapter, we introduced the concept of Explainable AI, a set of techniques that can be applied to ML models, after they have been built, to explain their

7 Roy McGrath et al., "Interpretable Credit Application Predictions with Counterfactual Explanations," arXiv, 2018, *https://arxiv.org/abs/1811.05245*.

8 Dr. Matt Turek, "Explainable Artificial Intelligence (XAI)," DARPA.

behavior for one or more predictions made by the model. We also explored why Explainable AI is needed by different groups who work with ML models, such as ML practitioners, observers, and end users. Each of these types of users has different needs for explainability, ranging from improving the quality of a model to building confidence in the model's effectiveness. To demonstrate this, we looked at several case studies of how explainability has been used. We started by contrasting two real-world examples in medicine, where one set of ML practitioners was trying to debug a poorly performing model used for classifying diseases from chest X-rays, while another group, ophthalmologists, needed to understand why their model had made a novel discovery about the inside of our eyes. To provide an introduction to other ways Explainable AI has been used, we also looked at other use cases across sales, fintech, and the defense industry. This introduction should help show you the variety of ways that explainability can be used, from different types of data and explanations, to the universal need for explainability regardless of the specific domain you are working in and the problem that ML is solving for your business.

In the rest of this book, we will discuss in more detail the tools and frameworks you need to effectively use Explainable AI as part of your day-to-day work in building and deploying ML models. We will also give you a background in explainability so you can reason about the trade-offs between different types of techniques and give you a guide to developing responsible, beneficial interactions with explainability for other ML users. Our toolbox covers the three most popular data modalities in ML: tabular, image, and text, with an additional survey of more advanced techniques, including example- and concept-based approaches to XAI and how to frame XAI for time-series models. Throughout this book, we try to give an opinionated perspective on which tools are best suited for different use cases, and why, so you can be more pragmatic in your choices about how to employ explainability.

In Chapter 2, we give you a framework for how different explainability methods can be categorized and evaluated, along with a taxonomy of how to describe who is ultimately using an explanation to help clarify the goals you will have in developing an Explainable AI for your ML model.