Deploying AI Systems Responsibly: Evaluating Explainable AI Techniques for Trustworthy Decision Support

Will Calandra

Georgetown University

BADM 399: Senior Thesis II

Professor Peter Jaworski

Professor Robin Dillon-Merrill

May 8, 2023

Table of Contents

Acknowledgements

I am extremely grateful to my faculty advisor, Professor Robin Dillon-Merrill, as this project would not have been possible without her mentorship and support throughout the entire process. I would also like to express my gratitude to Professor Peter Jaworski for his dedication in leading the Senior Honors Thesis class and creating a close-knit academic environment. Additionally, I would like to thank Professor Francesco D'Acunto and Professor Timothy DeStefano for serving on my defense committee and providing their guidance and expertise to inform my research. Lastly, I had the pleasure of working with a group of intelligent and cordial classmates who made this endeavor such a fulfilling experience.

Abstract

This research explains why the results of unexplainable artificial intelligence (AI) algorithms, while empirically accurate, should not be trusted to make decisions at scale, especially in risky situations. In particular, this paper will cite the "black box" characteristic of unexplainable AI algorithms as a critical bottleneck for AI system adoption and evaluate explainable AI systems as solutions to current algorithmic trust and safety issues. An overview and technical analysis of explainable AI methodologies will illustrate both benefits and challenges to transforming black boxes into interpretable decision support systems. A literature review of the psychology behind explanation will validate the current trajectory of explainable AI research as an algorithmic safeguard and highlight opportunities for future methodologies to address. Further, an analysis of explainable AI system use case successes and failures will identify and situate beneficial applications for explainable AI procedures to be adopted for practical use. This analysis will substantiate a need for interpretability in the design process, where a human can apply both a common sense and ethical framework to an AI system to mitigate unintended consequences and satisfy at-risk stakeholders. A behavioral study will then simulate factors of risk and explainability to decision makers to determine the appropriate applications where explainable AI systems can improve decision-making performance, trust, and safety. This research seeks to inform the development of a framework to design and monitor AI systems that leverage explainable AI's compatibility with making systems that people will adopt and trust.

Keywords: Explainable AI, Interpretable AI, Algorithmic Aversion, Algorithmic Safety and Trust

Introduction

Recent optimism surrounding AI seems justified: given large volumes of data, the capability of computers to process information has been proven to exceed human cognitive abilities (Anderson et al., 2018, p. 4-5). Moreover, the revolutionary processes in which a computer can autonomously learn and infer phenomena leads to the perception of AI systems as applicable to solve many problems and augment human capabilities. Such promises make it desirable for people to entrust their decision-making processes to AI systems, with their proven performance on training data being enough evidence for people to buy into AI algorithms as making optimal and informed decisions (Anderson et al., 2018, p. 2). Consequently, in businesses, governments, and scientific affairs, AI is starting to cross the chasm into becoming a central part of the decision-making process for key stakeholders. As such, interest in AI as a breakthrough information technology has led to increasingly ambitious use cases of AI to answer society's most challenging and complex questions (Mittal et al., 2022, p. 29-43).

However, the picture becomes less rosy given the limitations of AI systems. While it is common knowledge that the scope of an AI system is both sensitive and restricted to its data, the lesser-known limitation that some AI systems lack explainability should be a key concern for decision makers when deploying an AI system for practical use. While popular statistical frameworks offer well-known equations to interpret and diagnose an algorithm's results, when given a series of complex, nonlinear equations, it is practically impossible for a human to interpret how an algorithm achieved its results. Consequently, many current applications involving AI algorithms are known as "black box" systems, which are not only beyond the interpretation of the end user of an AI, but even the engineers who build such systems (Sarker, 2021, p. 410). Therefore, while AI systems may render accurate results in terms of fitting their data, the decision-making processes of AI systems are often inaccessible, which induces risk for decision makers when AI algorithms are deployed in real-world scenarios (Ribiero et al., 2016, p. 1).

One developing subfield of AI that attempts to solve this problem is called "explainable AI," where people are building algorithms and frameworks to trace the predictions made from an AI algorithm. Explainable AI can be defined as "a set of processes and methods that allows human users to comprehend

and trust the results and output created by machine learning algorithms" (IBM, 2022, p.1). Explainable AI has the potential to solve current algorithmic trust and safety issues involving black box algorithms, and this paper seeks to test explainable AI's ability to answer this call, especially in problem spaces involving high social costs and risk.

General key terms

Before providing an overview of explainable AI methodologies, it is important to define key terms and relationships foundational to the field of AI. The definitions below will help provide context to the terms used in this paper.

Artificial Intelligence: as defined in Janiesch et al. (2021), artificial intelligence (AI) is "any technique that enables computers to mimic human behavior and reproduce or excel over human decision-making," with the goal "to solve complex tasks independently or with minimal human intervention" (Janiesch et al., 2021, p. 686). AI is a broad term that captures this process, and an "AI system" describes an application of AI that is programmed in a computer, which often includes an actionable output or recommendation that completes a task.

Machine Learning: a subfield of AI that describes the more specific process by which a computer can iteratively capture patterns from data without being explicitly programmed. However, a human programmer still remains involved with representing the data inputs (features) to a machine learning algorithm, which is known as "feature engineering" (Janiesch et al., 2021, p. 686). Machine learning algorithms are "trained" on a partition of available data, "tested" for performance on a hidden partition of available data, and then "validated" on further hidden partitions of available data.

Deep Learning: a subfield of machine learning that is based on complex neural network algorithms, which are meant to simulate the architecture of the human brain, forming connections and discovering patterns in an autonomous way. Deep learning often deploys nonlinear equations, called activation functions, to represent the features from data. Therefore, in deep learning, no human input is

required, which is a differentiating factor compared to “shallow” machine learning methods (Janiesch et al., 2021, p. 687-689).

Engineer: a person who uses their technical skills to build an AI system. This person can be a statistician, computer scientist, data scientist, machine learning engineer, or any person involved in similar technical roles who build AI systems.

Decision Maker: a person tasked with making a choice based on an AI system’s output. This definition is interchangeable with an “end user” or “user” of an AI system, as this person is the final authority who acts upon an AI system’s recommendation.

I. Taxonomy of Explainable AI Methodologies

This section provides an overview of the motivations behind explainable AI, its key terms, common frameworks and approaches, and current technical limitations.

Motivations behind explainable AI

In order to provide a broad description for the motivations behind explainable AI, it first helps to understand the stages of a typical process for building an AI system. Often, an engineer follows the protocol of processing data inputs (including sourcing data, checking its quality, handling missing values, etc.), choosing an AI algorithm to best fit the data and validate its predictions (the training/testing/validation modeling process of statistical equations), and then deploying the system in practice, monitoring the input stream of data and the subsequent performance of the model’s predictions (Janiesch et al., 2021, p. 688-691). Such a task requires a deep understanding of mathematics, computation, and domain knowledge, so that in use cases of poor or nonsensical predictions, an engineer would be able to detect that a model is underperforming and debug its mistakes.

A new wave of complex machine learning frameworks, particularly in the domain of deep learning, has brought models that can involve millions of parameters and complex, nonlinear equations to the mainstream. According to a report by McKinsey & Company (2022) on the current state of AI, 30%

of companies surveyed stated that they have deployed deep learning models in at least one function or business unit (McKinsey & Company, 2022, p. 3). A trend towards deploying deep learning algorithms makes sense given its superior performance to shallow machine learning algorithms in use cases of high dimensional data (Janiesch et al., 2021, p. 688). However, with deep learning algorithms' complex equations, size, and abstractions on features, such modeling choices remove a level of understanding and accessibility for engineers to their models' internal processes. As such, the modeling procedure becomes reduced to a black box, as engineers optimize for the model's evaluation metrics (accuracy, precision, recall, etc.), thus offering little insight into the model's causal and correlated relationships. The downstream effects of this information loss have widened a gap between the research community and business sectors for deploying AI technologies, especially for business sectors that are heavily regulated (Arrieta et al., 2019, p. 86). Key stakeholders are not convinced solely by a model's performance in an experiment but seek to understand its fidelity for making correct and responsible decisions "in the wild" (Ribiero et al., 2016, p. 1). Compared to available data, real-world data is not known a priori, and can be significantly different from what was used to train a machine learning model, resulting in a problem that has been named "data set drift." In a model with millions of parameters, this effect can exacerbate to the point where the model's evaluation metrics are not indicative of its future performance, resulting in the need for more granular information (Ribiero et al., 2016, p. 1).

Research has confirmed that decision makers are often dissatisfied when only presented with a recommendation given by an AI, because *explanations* are desired to justify a model's predictions. In cases of unexpected or unfamiliar outputs, explanations become critical to measuring an AI's capacity to make informed decisions (Liao et al., 2021, p. 5). Furthermore, in high-stakes situations, such as acute medical diagnoses, humans with valuable domain knowledge need to be able to understand an AI's recommendations so they can complement the system to ensure task success (Zhang et al., 2020, p. 1). Therefore, since explainable AI is aimed at making black box models interpretable, it becomes a desirable design principle for a few reasons: to detect sources of bias for making adjustments to training data, gauge a model's robustness to understand how mutations can change predictions, and ensure that only variables

relevant and meaningful to the task predict the outputs (Arrieta et al., 2019, p. 83). This outcome can be achieved in a multitude of ways, a few of which will be discussed in the next section of this paper. Explainable AI is concerned with explaining how a model obtains its outputs, but it can be difficult to obtain an objective metric that measures the quality of such explanations or how interpretable an explanation might be to an end user (Arrieta et al., 2019, p. 101). Nevertheless, engineers can improve interpretability by reducing a model's parameters for digestible insights, visualizing the inner workings of a model, or they can even use natural language outputs to support a model's predictions (Arrieta et al., 2019, p. 85). However, it is important to note that in use cases that require a certain level of complexity, there is a trade-off between interpretability and model performance (Arrieta et al., 2019, p. 100). As discussed later in this paper, strong desires for explanations paired with explanations' capabilities to safeguard against subtle, hard to detect algorithmic errors can outweigh any decreases in model performance and may even lead to better model performance.

Explainable AI: key terms

In order to evaluate explainable AI's objectives, it is critical to first define the scope of the otherwise ambiguous terms associated with the field. Different papers in explainable AI often use the below terms interchangeably, so it is important to clarify their definitions for the context of this paper.

Understandability: a characteristic of a model in which a human understands how it works without any need to explain its internal structure or calculations (Arrieta et al., 2019, p. 84).

Interpretability: a characteristic of a model referring to the degree to which "a given model makes sense for a human observer" (Arrieta et al., 2019, p. 84). An interpretable model is one in which a human can understand the entirety of its inputs, calculations, and outputs, and can therefore be explained to a decision maker. It is the goal for explainable AI methods to increase models' interpretability (Arrieta et al., 2019, p. 88). Interpretability can also be referred to as transparency (Arrieta et al., 2019, p. 84).

Explainability: Arrieta et al. (2019) define explainability in the modeling sense as "the details and reasons a model gives to make its functioning clear and easy to understand" (Arrieta et al., 2019, p.

85). Literature from both explainable AI and the social sciences agree that it is difficult to reach consensus on a definition for explainability, but "good" explanations are known to be digestible, accurate, and fit for an audience.

Post-hoc explainability: the ability for a model to transform an uninterpretable model into an interpretable one, but only *after* the uninterpretable model has been trained. Post-hoc explainability is both a common and intuitive method for making uninterpretable algorithms interpretable, so this type of explainability will be discussed throughout the rest of this paper (Arrieta et al., 2019, p. 88).

Evaluation criteria for interpretability

Since the goal of explainable AI is to increase a model's interpretability, and it is difficult to measure interpretability, Arrieta et al. (2019) offer a framework to evaluate a model's interpretability by three levels, or degrees of interpretability: algorithmic transparency, decomposability, and simulatability. If a model satisfies any of the three levels, then the model can be considered interpretable, where an explainable AI is either sufficient or not needed (Arrieta et al., 2019, p. 86-90). These three levels are defined as follows:

Algorithmic transparency: the ability of a human to understand the process of a model for producing any given output from its input data. A satisfactory example is that the equation and error space of a linear model can be reasoned about and understood. A transparent model must be fully explorable by mathematical methods, so that it can be interpreted and thus understood (Arrieta et al., 2019, p. 88).

Decomposability: the ability to explain all parts of a model: its inputs, calculations, and outputs. Decomposability is satisfied when a human can explain every part of the model without the need for additional tools. A model that is decomposable is algorithmically transparent, where the differentiating factor is that no further mathematical exploration is needed to explain a decomposable model (Arrieta et al., 2019, p. 87-88).

Simulatability: the ability of a model to be "simulated or thought about strictly by a human" (Arrieta et al., 2019, p. 87). Simulatability can be achieved in a few ways; for example, one can visualize

a model's components to communicate its inner workings to someone else. A key consideration is that a model must remain simple enough that a human can understand the model in its entirety. A model that is simulatable is also decomposable and algorithmically transparent. In this case, one can understand the system "as a whole," which is its differentiating factor from decomposability (Arrieta et al., 2019, p. 87-88).

Explainable AI approaches

In order to provide interpretability to uninterpretable models, explainable AI can deploy post-hoc explainability methodologies, which can be implemented using three types of schemes: 1) model agnostic explainability methods, which can be applied to any model type, 2) shallow or deep machine learning explainers, which are tailored for a specific machine learning model, or 3) a hybrid scheme between the two. This paper is primarily concerned with model agnostic methodologies, because in studies where data scientists (engineers) were granted the freedom to choose among multiple explainability methods, they showed a tendency to use the same method (Krishna et al., 2022, p. 14). If data scientists repeatedly use the same method to explain their models, then it can be reasoned that model agnostic procedures are more likely to be used in practice, assuming that data scientists are likely to deploy multiple model types in their workflow. Therefore, while narrow, specific explainability methods may be useful for niche applications to boost performance, they are outside of the scope of this paper. A few key terms for model agnostic, post-hoc explainability methods are listed below:

Explanation by simplification: a type of explanation that decomposes a larger model through approximation by a simpler model. This technique also includes approximations on subsets of data, particularly at the example level (these are called local explanations**)**. For example, the LIME method, or local interpretable model-agnostic explanations, builds locally linear models around the predictions of a black box model, thus using an interpretable (linear) model to provide explanations on smaller, simpler subsets of data (Arrieta et al., 2019, p. 92).

Explanation by feature relevance: these explanations describe a black box model by ranking the influence of each feature on its predictions. An example of a framework supporting this explanation would be the SHAP method, or Shapely Additive Explanations, which aggregates feature importance scores for each prediction that sum to the probability of the predicted class (Arrieta et al., 2019, p. 92).

Explanation by visuals: these explanations are typically combined with simplification or feature relevance explanations. In the context of LIME and SHAP, a visual explanation from LIME could involve displaying the pixels of an image that had the highest effect on an example's prediction, while a visual explanation from SHAP could involve a feature importance plot, which would indicate the ranking and direction of a feature's influence on a model's predictions (Arrieta et al., 2019, p. 92).

Explanation by counterfactual: These explanations tweak a model's inputs and test a threshold for changing a model's predictions. This technique explains the sensitivity of a model and is often used to assess robustness and quality of predictions for similar cases (Arrieta et al., 2019, p. 101). For example, a LIME implementation can perturb instances of an image, and run each instance to get a probability of prediction, thus validating a model's robustness. A display of the most relevant pixels of an image would then explain its prediction (Arrieta et al., 2019, p. 97).

Technical limitations to explainable AI

As previously mentioned, a key point of tension with explainable AI approaches is that there appears to be a trade-off between interpretability and model performance, even if post-hoc explainability methods are deployed (Arrieta et al., 2019, p. 99-100). Models that are more complicated, to the point where they are beyond the grasp of humans, are more likely to be accurate, because their functions are more flexible and can serve as a better fit for the data. Explainable AI can set out in its research to improve this trade-off, but it is important to note that its explanations (and therefore candidate model classes) are limited by the capacity of its audience to comprehend its processes (Arrieta et al., 2019, p. 100). Furthermore, as previously mentioned, it is difficult to capture the quality of a model's explanations, as there is no unified, industry-standard metric to evaluate what constitutes a "good"

explanation (Liao et al., 2021, p. 1). The fact that explanations are relative to audiences also implies that the same explanation will vary in quality across individuals. One can also imagine explanations as relative to the knowledge that an individual seeks from an explainable AI system (Liao et al., 2021, p. 2). For example, Liao et al. (2021) conducted a survey of User Experience (UX) design practitioners who were concerned with the static nature of current explainable AI methods. When presented with an explanation of a particular example, respondents complained that they were unable to ask follow-up and clarifying questions to an explainable AI. In this way, some people may not be satisfied with the seemingly surface-level explanations provided from an explainable AI, which indicates a rigidity in explainable AI techniques that should be remedied for wider adoption (Liao et al., 2021, p. 5).

Another technical concern about explainable AI is that using multiple methods can introduce what Krishna et al. (2022) call the "disagreement problem," where different post-hoc explainability methods draw divergent conclusions about an AI's predictions. In current research, a playbook to resolve the disagreement problem in the model evaluation process does not exist. In a study of data scientists, 84% of interview participants stated that they encountered a disagreement problem in their daily workflow, and 86% did not know how to solve the issue. Such problems came from different orderings of top features or changes in sign of feature importance between methods. This result is particularly concerning for complex models, because disagreement among explainable AI can increase with greater model complexity (Krishna et al., 2022, p. 3). In particular, LIME and SHAP have been shown to produce not only inconsistent but unstable explanations in such cases (Krishna et al., 2022, p. 2). As such, the utility of explainable AI algorithms to make models interpretable might have diminishing returns for more complex models, which makes sense given that explainable AI methods like LIME only use simple mathematical approximations across localities of data. Furthermore, Krishna et al. (2022) mention that in the case of contrastive explanations where perturbations are known, models can be exposed to adversarial attacks and "fair washing," where an explainable AI model can be manipulated to serve a malicious end (Krishna et al., 2022, p. 2). In this way, an explainable model can easily be tinkered to serve the biases and malpractices that it might set out to solve in the first place.

II. Psychological Considerations of Explainable AI

This section evaluates the psychological factors applicable to explainable AI methodologies to highlight both the field's progress and focus areas for improvement.

Psychological justifications in favor of explainable AI

Successful deployment of an AI system requires an accurate mental model of both the task for an AI to solve and its processes for completing that task. The formation of a mental model for an algorithm to be understood requires it to be interpretable, which is why the development of successful explainable AI methods are critical for improving human-machine interactions. For example, Zhang et al. (2020) define a successful mental model for an AI to be a correct mental model of the AI's error boundaries. When an AI system is used for decision support, the human's understanding of a model's tendencies for error ultimately decides whether trust is awarded to an algorithm to make a decision (Zhang et al., 2020, p. 1). In order to identify situations where a model is likely to make an error, the user should focus on understanding specific cases and examples in which an algorithm will be prone to mistakes. Therefore, local explanations become preferable to global explanations from an explainable AI system. Additionally, explainable AI algorithms can assist a user's calibration of an AI system's capabilities, where they can learn its strengths and weaknesses from its explainable and decomposable components (Zhang et al., 2020, p. 1). While repeatedly interacting with explanations may seem like a tedious exercise, Miller (2018) contends that if explanations are beneficial to both learning and generalization, then individuals will require less explanation as they continue to interact with a system, thus developing a level of simulatability with the system (Miller, 2018, p. 47). For this reason, a user's mental model and understanding of an algorithm improves over time, which means that in cases of model errors, there will be fewer surprises for the user.

An improved understanding of an algorithm can also be demonstrated through a guided debugging process, where users can interact with an AI system and make changes to improve its performance. In an experiment to validate the LIME method, Ribiero et al. (2016) proved that non-experts

improved a model's performance to generalize through feature engineering after they were presented feature relevance visualizations. In this way, the respondents were able to easily remove the "faulty" features that were present in the simulated model by simply viewing the feature relevance explanations. Furthermore, even when provided a faulty data set, respondents were able to pick a better classifier for generalization when given the feature relevance visualizations, which provides evidence that when given explanations, even non-experts can form a strong understanding of an algorithm and its flaws (Ribiero et al., 2016, p. 8).

Additionally, explanations are more intuitive to humans than probabilities and percentages. For example, Miller (2018) claims that mentioning probabilities are not as effective as mentioning *causes* when providing an explanation. Miller (2018) also states that while statistics present the most likely outcome for an event, the explanation for the "most likely" case may not correspond with the best explanation for a person; however, if an underlying cause is presented for a statistical generalization, it can create more satisfaction towards an algorithm (Miller, 2018, p. 6). Therefore, the ability for explainable AI methods to provide reasons for specific, example-based predictions produces more satisfying and comprehensive analyses compared to unexplainable methods.

Explanations also offer strong logical support for algorithms' most curious cases, where an output may seem unexpected or nonsensical. In the UX survey conducted by Liao et al. (2021), understanding particular decisions was ranked by designers as a top reason for desiring explainability. The desire to understand a prediction was naturally triggered after surprising and abnormal events, which is summarized by a respondent as such: "for everyday interactions, most likely it's how did the system give me this answer? Not just any answer, but all of a sudden, here's this thing I'm seeing" (Liao et al., 2021, p. 8). Miller (2018) supports this theory by characterizing explanations as contrastive, where there must be some event to which a comparison is made. For example, Miller (2018) states that "people do not ask why event P happened, but rather why event P happened *instead* of some event Q" (Miller, 2018, p. 6). This attitude aligns with the capability of explainable AI to provide counterfactual explanations, where example predictions are tested against some benchmark. In this way, counterfactual explanations can start

to address explainable AI's need to cater to an audience, because social expectations can be set as the default benchmark example in a contrastive explanation. While some may challenge these choices as potential sources of bias, Miller (2018) claims that selection bias is not a barrier to understanding an AI system. Miller (2018) supports this statement by claiming that people are not interested in the cognitive burden of a complete explanation and cause for an event, but rather accept one or two causes as a satisfactory explanation. Therefore, explainable AI systems should not be concerned with providing a complete logical explanation of predictions, but they should instead elucidate the causes that are important to the outcomes by providing explanations of specific examples (Miller, 2018, p. 6).

Psychological evidence against explainable AI

While there is plenty of evidence in support of explainable AI's utility to improve one's understanding of an algorithm, there exist a few hurdles that the explainable AI community must address. First, as previously mentioned, explainable AI offers static recommendations, where any further line of questioning is often incompatible with the current technical capabilities of explainable AI. For example, Liao et al. (2021) cites a user disappointed with the comprehensiveness of explanations, who stated "explainability isn't just telling me how you get there, but also, can you expand on what you just told me" (Liao et al., 2021, p. 5). Both Liao et al. (2021) and Miller (2018) argue for explainable AI to cater to a line of questioning, which means that future frameworks should be developed for explainable AI to explore and debate the reasoning behind its explanations (Liao et al., 2021, p. 5; Miller, 2018, p. 8). Furthermore, Miller (2018) characterizes explanations as social, which requires both conversation and interaction. As such, when dealing with a diverse set of individuals, an explanation therefore becomes contextual, where only a few causes and reasons for an event actually matter to a particular individual (Miller, 2018, p. 6-7). While local explanations, feature relevance, and contrastive explanation techniques play into this idea, explanations must have a proxy of pre-existing beliefs to anticipate a user's questions and offer sensical explanations. This capability was not found in current post-hoc explainability methods.

Furthermore, communicating explanations can fail if the engineer either overloads users with information or fails to present explanations in a persuasive manner. While explanations are meant to be a support mechanism for making informed decisions, studies cited by Zhang et al. (2020) suggest that overloading users with information may lead to decreased performance in AI-assisted decision making (Zhang et al., 2020, p. 9). Miller (2018) summarizes a framework from Hilton to try to identify a balance between useful information and overload, in that explanations should "only say what you believe, only say as much as is necessary, only say what is relevant, and say it in a nice way" (Miller, 2018, p. 11). In the case of explainable AI, the first three criteria seem satisfied with approximations, example-based explanations, and feature relevance. However, depending on the use case, an explainable AI needs to be engineered to maintain a balance between useful explanation and information overload, which is difficult to quantify for every user. Miller's (2018) "say it in a nice way" principle is also a roadblock, especially if many outputs from explainable AI are not given in natural language. "Say it in a nice way" also implies a level of persuasion, which explainable AI attempts to do in its outputs, but only through the logic in which those outputs are communicated. Future studies should examine the capabilities of explainable AI to persuade in its explanations while maintaining its fidelity to both the model and task at hand. Information overload and persuasion are key pillars to effective explanations but are areas that current explainable AI methods still leave in the hands of its engineers.

III. Use Case Analysis

This section introduces milestone use cases that illustrate both the benefits and drawbacks to deploying explainable AI in the real world.

Cases beneficial to explainable AI

First, use cases that substantiate support for explainable AI are those involving data leakage. Data leakage refers to cases in which a signal that would not appear in a model's real-world application "leaks" into the model's training data, and thus its calculations. Data leakage is different from data set drift, as

data leakage involves nonsensical features, while data set drift involves poorly estimated distributions for new data. Explainable AI helps to easily detect data leakage through its feature relevance capability. As an example, in a medical application, a feature relevance chart was able to detect that a black box AI system was using the feature "Patient ID" to predict a target class (Ribiero et al., 2016, p. 2). Patient IDs, while accessible, are unique identifiers that are not representative features in real-world testing data and would therefore be poor predictors. In this case, it would be difficult for an unexplainable model to detect the correlations between a patient's ID and a target class. However, feature relevance explanations can easily alert engineers that an algorithm, while properly trained using a black box model approach, was making predictions for the wrong reasons. Now suppose that the aforementioned unexplainable model was to pass the initial checks, and therefore be deployed to generalize and make recommendations at scale. It is obvious that such an unchecked risk is why, as previously mentioned, sectors that are heavily regulated and operate in high-risk situations are hesitant to approve black box AI systems for practical use (Arrieta et al., 2019, p. 86).

Humans can also apply their domain knowledge to an algorithm's explainable components, subjectively validating its processes for making its predictions. For example, in an imaging data set of guitars in which a deep learning algorithm was applied, the superpixels generated from visual and counterfactual explanations showed that a guitar's fretboard was the most substantial predictor for misclassifying an acoustic guitar as an electric guitar. Even though the example's prediction was *wrong*, it proves that the algorithm was "not acting in an unreasonable manner," because people know that the fretboards for both guitar types are very similar, which would explain the algorithm's error space (Ribiero et al., 2016, p. 5). Therefore, in this case, a user can trust the algorithm to behave appropriately, and even though it will make errors, its potential to generalize is much greater than that of a comparable, nonsensical algorithm. Additionally, upon observing this phenomenon, the user would have the foresight to predict the cases in which the model is likely to misclassify a guitar. As previously mentioned, a guided feature engineering and debugging process would likely decrease model errors, or at least eventually prove to its user that the AI system is not well-suited for the desired task.

Cases against explainable AI

A first use case in which explainable AI struggles is when deeper and detailed explanations are required by its audience. For example, a LIME algorithm can be deployed to diagnose patients with the flu, returning symptoms such as "sneeze" and "headache" as the top explainers for the AI system's diagnosis. However, this explanation may not satisfy a doctor's needs, who may require a further line of questioning to the algorithm to understand the logic in its prediction (Ribiero et al., 2016, p. 2). As previously mentioned, such a capability eludes current explainable AI systems (Liao et al., 2021, p. 1).

Furthermore, use cases in which non-experts have similar levels of domain knowledge as an AI system caused explanations to become superfluous. For example, in a study conducted by Zhang et al. (2020), participants were tasked with predicting the income level of a person based on demographic and job characteristics. The participants, who were non-experts in the task at hand, were briefed with information that was similar to what was provided to an AI system. In the study, simply displaying a model's confidence score improved participants' trust and willingness to rely on an algorithm, especially in cases of high algorithmic confidence (Zhang et al., 2020, p. 6). In contrast, when local explanations were displayed, the accuracy of participants' decision making decreased across all algorithmic confidence levels (low, medium, and high confidence). This result may be due to information overload, or the fact that participants did not have an accurate mental model of success for the task at hand. Zhang et al. (2020) acknowledged in the study that they did not measure the mental models of participants; however, one can conclude from this study that in cases of similar domain knowledge, the error boundaries are more likely to align between a human and an AI system (Zhang et al., 2020, p. 8). Therefore, there is little benefit for an AI to explain its predictions because a human is already unable to understand and form their own mental model of the task to be completed.

IV. Methodology

In order to evaluate the effectiveness of explainable AI methodologies, a 2 x 2 factorial behavioral survey design was conducted, varying the factors of use case and explainability in a simulated,

AI-assisted decision-making environment. One use case, or "scenario" chosen involved decisions related to the hiring process for a fictitious managerial job, which was intended to be a high-risk scenario. In contrast, the other scenario involved decisions to set baseball card prices at a fictitious trade show, which was intended to be a low-risk exercise. Participants were Georgetown University students who received class credit for their completion of the task as part of a school subject pool. Students were randomly assigned to one scenario and one explainability procedure (either shown model explanations or given black box model outputs throughout the entire simulation of one scenario). Each scenario is described below in more detail:

Scenario I: Hiring Algorithm

Hiring algorithms that use computer vision technology and machine learning to create an "employability" score have been criticized as a dangerous application of AI-assisted decision making. For example, HireVue's video interviewing system has been challenged to be a "license to discriminate," as the algorithms are trained to seek specific features in candidates' facial expressions, mannerisms, and tone of voice (Harwell, 2019). In this experiment, simulating a hiring algorithm was intended to be a "high-risk" situation in which a participant was tasked with making a hiring decision for four different candidates. The "risk" was elevated by raising the stakes of failure in the task's introduction (see Exhibit 1) paired with the "human element" that comes with evaluating other individuals, which creates social costs and the potential for bias in the decision-making process.

In order to help participants gain an understanding of the algorithm's task to calibrate their own thought process towards making hiring decisions, participants were provided with the following criteria in which to make their decisions: candidates were supposed to be smiling in the image shown, they were to mention specific keywords (or synonyms), and have 5 or more years of experience (see Exhibit 2). All participants were then asked to make decisions for hiring four candidates, where they were presented with an image of the candidate (to detect their smile), a summary of keywords used (to detect similarities with company objectives), and the candidate's years of experience (to detect fit with the experience criterion).

To isolate the effects of explainable AI methods in the decision-making process, participants assigned Scenario I were split into two groups: for each candidate, one group was shown model explanations, while the other group was not shown model explanations. Groups are referred to as either "the explainable" participant group or "the unexplainable" participant group in the next section of the paper. Model explanations consisted of simulating the LIME method on the candidate image for visual and counterfactual explanations, which were used to detect a smile by highlighting the most relevant pixels of the image. The image explanations were called "image highlighting" to participants in order to make the term "explanation by counterfactual" accessible to participants in the survey. Additionally, the SHAP method was simulated for feature relevance explanations, which provided the direction and effect of each feature on the model's candidate score, which determined the model's hiring recommendation (see Exhibit 3). Throughout the simulation, the feature relevance explanations were called "feature influence" to participants. Since the feature relevance explanations push each prediction in a particular direction, it is reasonable that "influence" would provide more semantic meaning than "relevance" to participants. In combination, both explanations were intended to provide enough information to satisfy the model interpretability level of decomposability, assuming the "algorithm" was already made transparent (or interpretable) from explanations. For each explainable AI method, a brief interpretation of the method's output was provided to the user for assistance. This choice was made to help AI non-experts interpret the explanations, satisfying the model's simulatability criterion.

Model explanations and interpretations offered more information than the group of participants who did not receive model explanations, or the unexplainable group (see Exhibit 4). Both groups of participants were given the model's confidence rating (which was "high" for all cases), the candidate score (or the model's estimated probability of a candidate being a good hire), the model's hiring recommendation (a binary yes/no decision), and an overall summary of all model outputs. Each participant was given a list of questions to answer after making each hiring decision (see Exhibit 5) in addition to reflection questions at the end of the survey (see Exhibit 6). The list of candidates presented is available in the Appendix (see Exhibit 7). Results and analysis will be discussed in the next section.

Scenario II: Baseball Card Pricing Algorithm

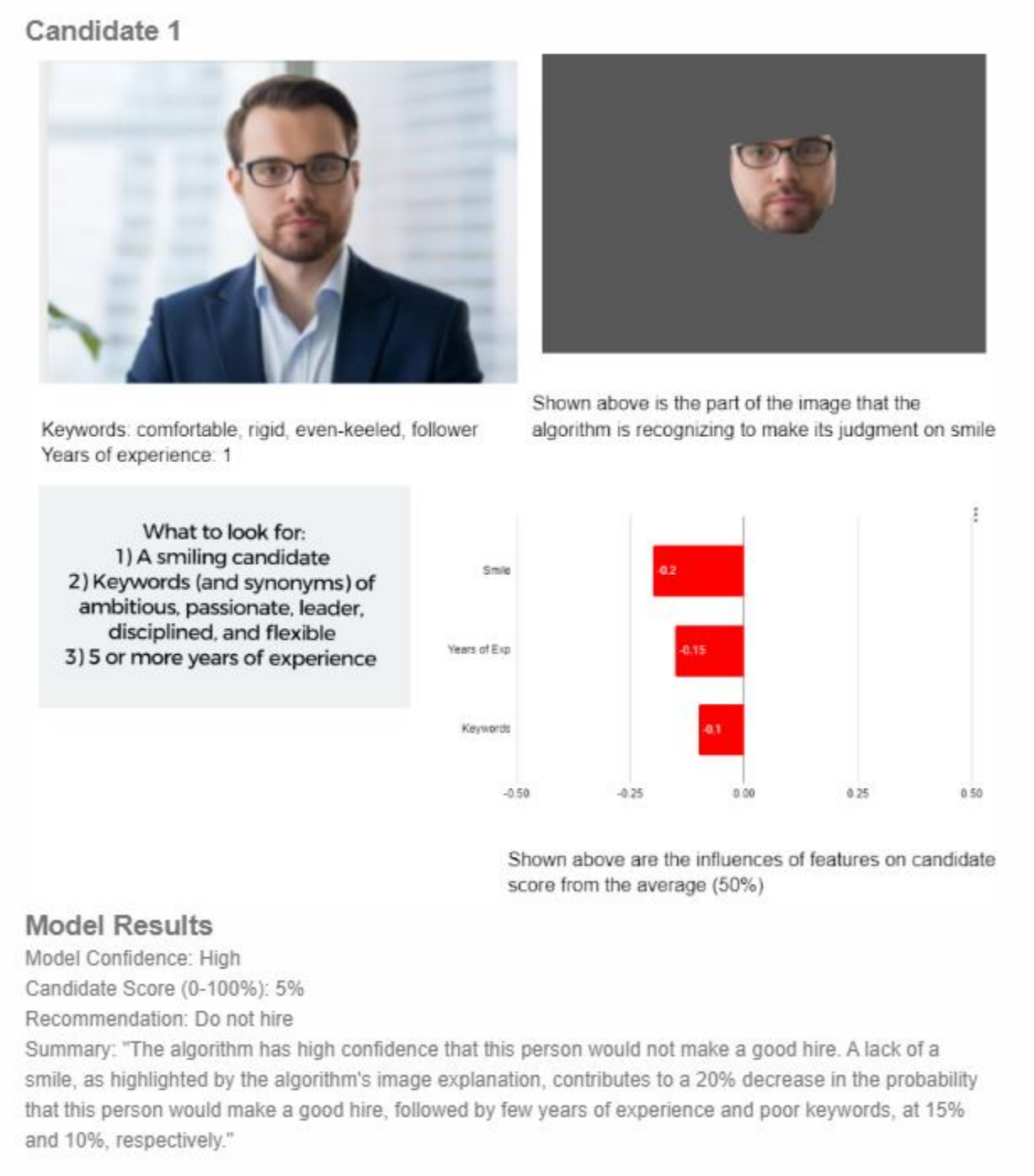
Baseball cards operate in a simple marketplace, where buyers and sellers agree upon prices at which a card is sold. Capturing the market value of a baseball card is mostly a quantitative process, where potential buyers and sellers can reference publicly available information to get an estimate on the valuation of a specific baseball card (Vince, 2022). In this experiment, participants were given the task of setting prices for four baseball cards, based on three established criteria: the age of the card, its condition, and the number of copies available on the market (see Exhibit 8). Scenario II is meant to replicate as close to a "purely quantitative," straightforward exercise as possible, where as previously mentioned from Zhang et al. (2020), model explanations would become superfluous and may contribute to information overload (Zhang et al., 2020, p. 9). As such, using an AI system for assistance in the decision-making process would be perceived as a "low risk" situation, and the stakes would be further lowered through the language used in the task introduction (see Exhibit 9). Since evaluating baseball cards does not involve the judgment of others, the "human element" from Scenario I is also eliminated.

The procedure for this experiment follows the same structure as Scenario I, as there were participants who were provided image, counterfactual, and feature relevance explanations (see Exhibit 10) and those who were not provided model explanations (see Exhibit 11). In both cases, participants were asked if they would recommend setting the price of the baseball card at the algorithm's output in addition to their confidence in both their decision and the algorithm's performance (see Exhibit 12). However, if participants did not agree with the price set by the algorithm, they were asked to set their own price, so the magnitude of the algorithm's effect on price changes could be tested. Nevertheless, just like Scenario I, participants answered a series of reflection questions at the end of the survey (see Exhibit 13). A complete list of the four baseball cards reviewed can be found in the Appendix (see Exhibit 14). Results and analysis will be discussed in the next section with some comparisons to Scenario I.

V. Results & Analysis

For this section, the results of each experiment will be discussed for both the hiring algorithm and baseball card scenarios. Since each experiment will be heavily referenced throughout this section, the explainable cases will be shown in sequence in this paper, and to stay concise, the unexplainable cases will be available in the Appendix. It is important to note that references to image highlighting and feature influences only correspond to explainable cases as they were not visible for unexplainable cases. Statistics comparing explainable and unexplainable cases will be reported, with their corresponding significance tests available in the Appendix. These trends will be combined with a qualitative analysis of open-ended responses, which will inform the conclusions that can be drawn from both cases.

Scenario I, Candidate I



Shown above is the first candidate for the hiring algorithm use case, which was intended to be a baseline case to give participants the opportunity to gain an understanding of the model explanations and offer a straightforward choice to not hire this candidate. The image highlighting shows a clear focus on

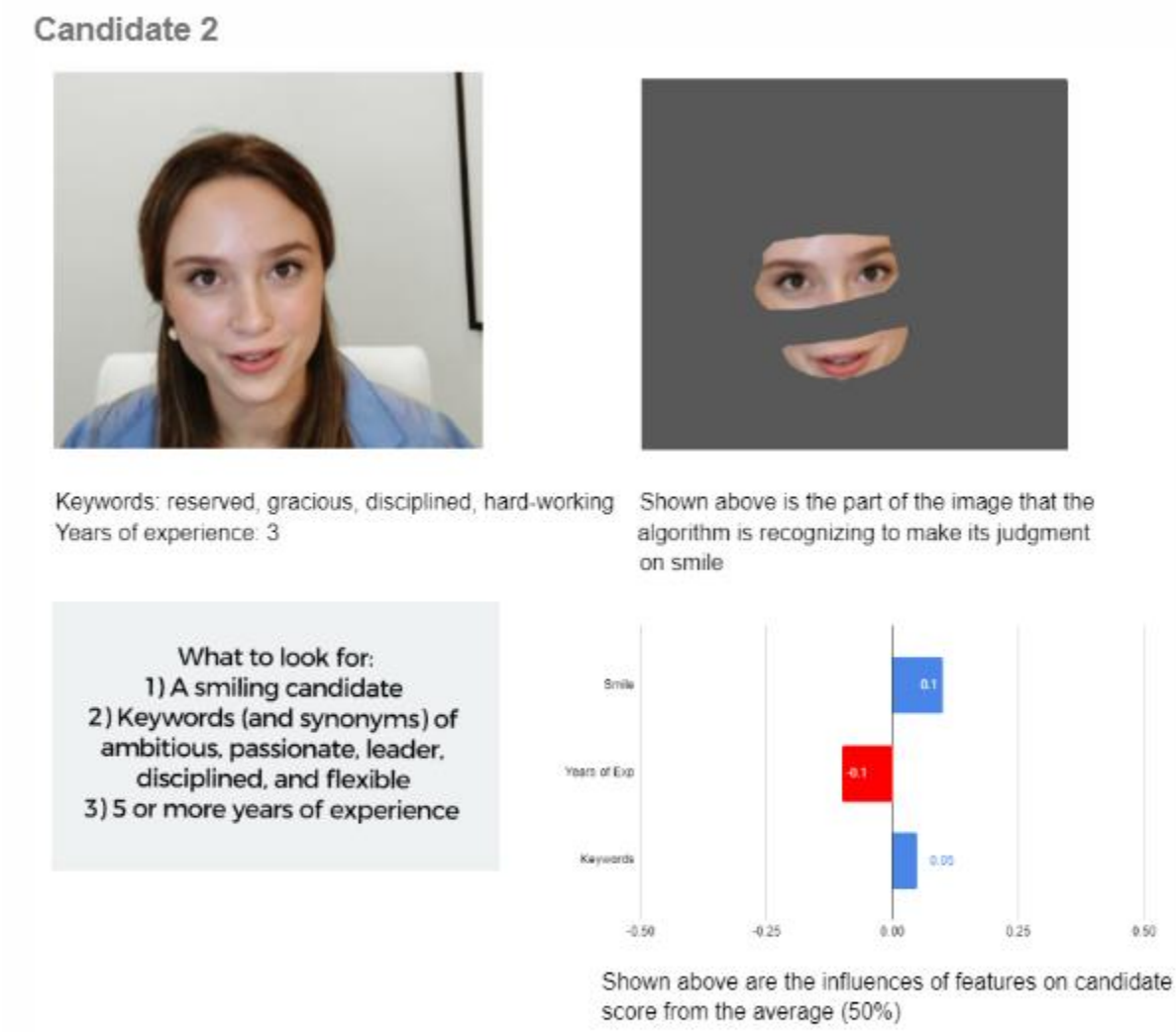
Candidate I's face with no smile, and the keywords of "comfortable, rigid, even-keeled" and "follower" are antonyms of the company's favored profile. Additionally, Candidate I possesses one year of experience, which is far below the five-year benchmark sought by the firm. Participants in the explainable group were supposed to notice that all three of these features would negatively impact the probability that Candidate I would make a good hire. The breakdown of the degree to which these features influenced the model's candidate score is shown at the bottom right of the model outputs.

Across both explainable and unexplainable cases, it was nearly unanimous from participants that Candidate I should not be hired, with 99% and 97% "do not hire" recommendations from both explainable and unexplainable cases, respectively (see Exhibit 15, A). However, even for a straightforward case, 58% of participants shown model explanations cite that they made their decision in balance with the algorithm, while only 38% of participants given an unexplainable model output balanced the algorithm along with their own decision. Across both groups, only 3% of participants default to the algorithm for making their recommendation (see Exhibit 15, B). Confidence scores were similar across both explainable and unexplainable participant groups in participants' decisions, the algorithm's recommendation, and its ability to generalize to make recommendations about more candidates. A surprising result was that participants' confidence ratings were higher for their own recommendations compared to the algorithm's recommendation, even though both decisions yielded the same result. The fact that participants were not as easily convinced from the baseline case that the hiring algorithm would generalize is an appropriate reservation, yet surprising to observe from a group of AI non-experts (see Exhibit 15, C).

From analyzing the results of Candidate I, participants are more likely to incorporate the hiring algorithm in their decision-making process, which is a testament to participants' early trust in this algorithm. There is clear agreement between both explainable and unexplainable participant groups that Candidate I should not be hired, which matched expectations. Of the explainable participant group, 65% of participants cited the feature influence bar chart as the most helpful for making recommendations,

which offers support for the strength of feature relevance and visual aids in the decision-making process (see Exhibit 15, D).

Scenario I, Candidate II

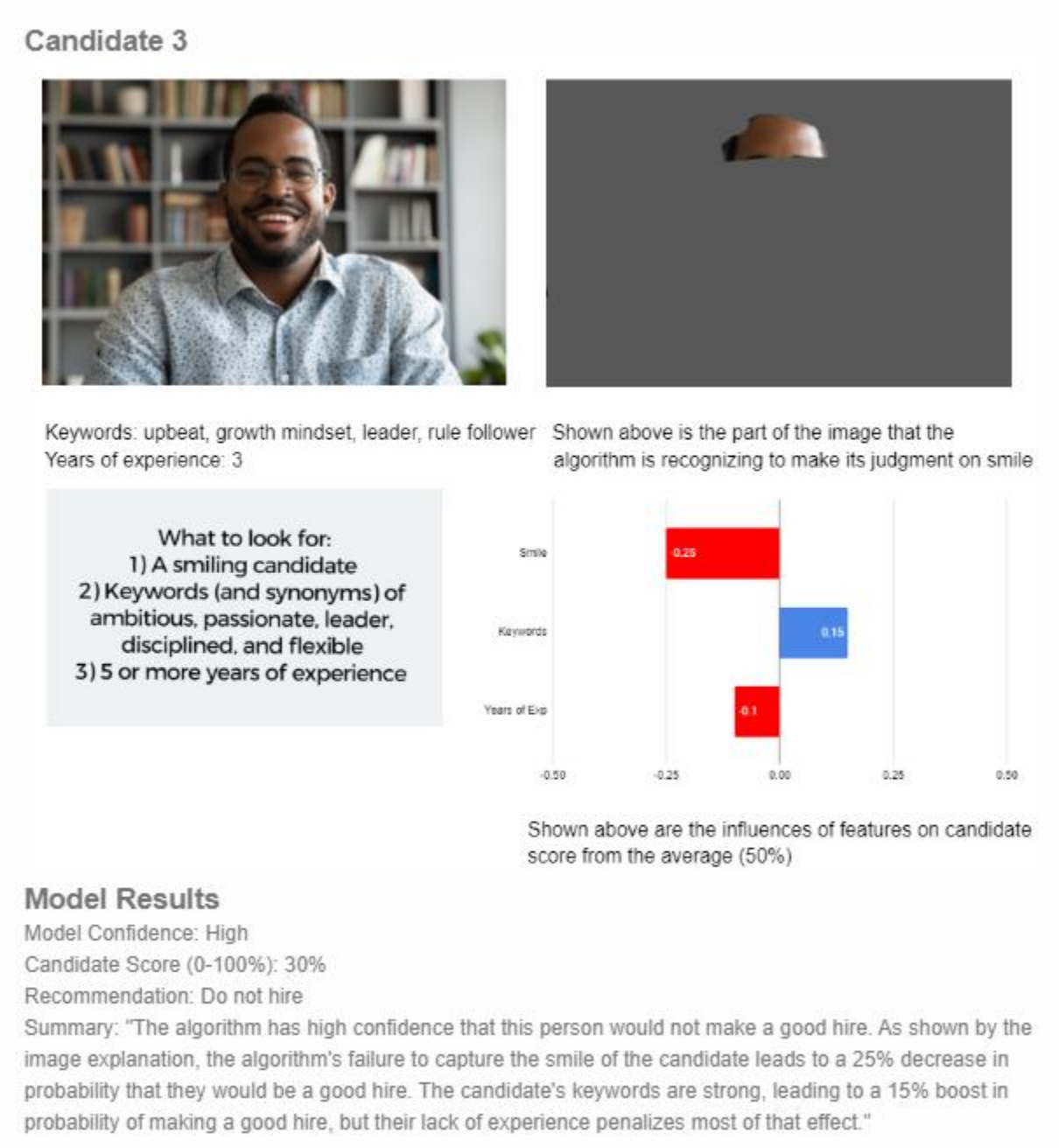


Candidate II was supposed to induce some confusion in the decision-making process, because the image of the candidate showed a slight smile that was not obvious to the naked eye. This calibration was intended to create an advantage for the explainable participant group in their recommendation process, because the image highlighting accurately identifies the eyes and mouth for the smile, which is then correctly reflected as a positive feature for "Smile" in the feature influence bar chart. Therefore, participants could piece together the appropriate image highlighting with the feature influence chart to understand why the model was at a 55% candidate score, which was above the 50% threshold, thus warranting a hiring recommendation. In the keywords, "disciplined" appears, which is a direct match with the company's profile. Additionally, three years of experience is below the desired five years of experience, but participants may recognize that three years is close to five, so the effect of inexperience should be perceived as small. Overall, participants were intended to be on the fence for Candidate II, but the smile, as accurately characterized by the model, was supposed to provide the proper insight to the explainable group for making the correct decision to hire this candidate.

While 55% of the explainable participant group elected to not hire Candidate II, this result is not significant enough to yield an overall hiring recommendation by the explainable participant group. In contrast, 65% of the unexplainable group chose against hiring Candidate II, which is a significant majority (see Exhibit 16, A). 75% of participants cited the feature influence as the most helpful tool in making their decision, which is likely because the bar chart revealed how the model balanced the factors in its hiring recommendation, which influenced the explainable participant group towards the correct hiring recommendation (see Exhibit 16, B). This observation substantiates the benefit of explainable AI for this case, because while both groups equally balanced their decisions with the algorithm, at 59% and 53% for the explainable and unexplainable participant groups respectively, the explainable participant group had better performance (see Exhibit 16, C). Confidence ratings across both explainable and unexplainable groups dropped relative to Candidate I, but were similar to each other, at 74 and 76 confidence in the recommendation, 57 and 53 confidence in the algorithm's recommendation, and 60 and 59 in the algorithm's ability to generalize, respectively (see Exhibit 16, D).

While the model explanations tried to tip participants in the right direction for interpreting Candidate II's smile, there was still confusion in how to evaluate their smile, as one participant stated, "one of the candidates didn't even look like she was smiling; it was just a picture of her mouth moving." Persistent confusion could be a potential reason why the explainable participant group reached an inconclusive result, but while both groups use the algorithm at a similar rate, Candidate II illustrates a case where the assistance of an explainable AI improves the performance of participants.

Scenario I, Candidate III



Candidate III was the clear case where an algorithm would return an unexpected result, which could be combined with participants' prior knowledge of algorithmic bias on minority groups. In this case, there is a clear risk of discrimination against minority groups given Candidate III has a clear smile and the proper keywords. Compared to Candidate II, Candidate III's traits are relatively strong, and both

candidates match on years of experience. Therefore, it was expected that participants would have a strong preference to override the algorithm's recommendation and choose to hire Candidate III. For the explainable group, the image highlighting clearly shows the algorithm's faults, as it is unable to capture the features of Candidate III's smile. For the direction of this effect, the feature influence chart reveals that the algorithm unfairly penalizes Candidate III by 25% on the smile, which is enough evidence for participants in the explainable group to reverse the algorithm's recommendation. Therefore, participants should make the correct recommendation by going against the algorithm's recommendation, choosing to hire Candidate III.

Both participant groups recommended hiring Candidate III, but there were significant differences across the groups' abilities to detect the algorithm's error. For example, 83% of the explainable group recommended hiring Candidate III, while only 69% of the unexplainable participants made a recommendation to hire (see Exhibit 17, A). Additionally, while a majority of both groups make their decision on their own, the explainable model participants correctly deviate from the model at a higher rate, with 80% making the decision on their own versus 70% of participants from the unexplainable participant group (see Exhibit 17, B). Furthermore, in a case where it would be irresponsible to use the algorithm's recommendations, confidence in the algorithm's recommendation for Candidate III was properly rated at 26 for the explainable participants, but the unexplainable participant group averaged a 46 rating. Additionally, in a diverse case of Candidate III relative to other candidates previously shown, since the model fails to make an appropriate recommendation, it should not be expected to generalize. Poor generalization is correctly identified by the explainable model group, as they rate their confidence in the model's ability to generalize at 45, while the unexplainable model group rates their confidence at a higher score of 54. Third, the explainable group expresses more confidence that they made the right recommendation, at a confidence rating of 83 compared to the unexplainable group's 77 (see Exhibit 17, C). However, compared to Candidate II, both groups experience significant decreases in confidence in the algorithm for making a decision, for both the current example and its ability to generalize (see Exhibit 17, D).

It is clear that the model explanations mitigate the risks of making hiring mistakes, as participants given explanations were more likely to override the algorithm and question the model's ability to make recommendations across a broad range of candidates.

Scenario I, Candidate IV

Given their previous exposure to a faulty imaging recognition software, the explainable participant group would be expected to have a better understanding of the model's capabilities and could look out for similar phenomena in subsequent examples, which is demonstrated with Candidate IV. Similar to Candidate II, the naked eye should observe a slight smile that would correspond to a positive effect on Candidate IV's score. However, the image highlighting shows that the algorithm grades Candidate IV's smile on the background of the image rather than their facial features. Therefore, paired with information from the feature influence chart, the explainable participant group would have the insight to reverse the algorithm's decision and make the correct recommendation to hire this candidate. Candidate IV also has very strong keywords that match the company's profile of "ambitious, disciplined," and "leader." However, just like Candidate I, Candidate IV has only one year of experience, which is penalized to the same degree as Candidate I. Nevertheless, the positive effect of the keywords and a smile should override the inexperience and lead to a hiring recommendation from participants.

Contrary to expectations, both groups recommend not hiring Candidate IV, but similar to Candidate II's case, Candidate IV receives more leniency, with 78% of the unexplainable participant group recommending to not hire Candidate IV, while only 64% of the explainable participant group make the same decision (see Exhibit 18, A). Additionally, in a case where the algorithm displays a clear fault, 67% of the explainable participant group make their decision on their own, while only 54% of the unexplainable group make their decision on their own. Therefore, the explainable candidate group is practicing more responsible use of the algorithm, refraining from its use at a higher rate than the unexplainable participant group in erroneous cases (see Exhibit 18, B). In fact, the biggest difference between groups for Candidate IV remains in the confidence levels in the algorithm, as the explainable group rates only 36 confidence in the algorithm's recommendation for this example and 41 for its ability to generalize. However, for the unexplainable group, the confidence rating in the algorithm remains at 61 confidence for the recommendation on Candidate IV and 55 confidence for the algorithm to generalize (see Exhibit 18, C).

In the case of Candidates III and IV, it is clear that it is more difficult for the algorithm to gain the trust of participants who were shown model explanations. Additionally, in both cases of Candidate III and IV, the model explanations point explainable participants in the right direction for hiring recommendations over the unexplainable participant group in addition to providing higher levels of decision-making confidence. This phenomenon is a direct result of viewing both the faults of the image highlighting and the effects of those errors on the model's predictions. It is important to note that in the decision-making process, the participants shown model explanations used the model to a lesser degree in cases of apparent failure, while the unexplainable participant group still balanced the model's inaccurate decisions at a higher rate.

Scenario I, End of Survey Results

At the end of the survey, the explainable participant group rated the algorithm as having a higher influence over their decision making, at a rating of 42 compared to a 35 rating by the unexplainable participant group (see Exhibit 19, B). However, both groups agreed that they had enough information from the model to make a decision with confidence. For the explainable participant group, there was high agreement for model explanations helping them better understand the capabilities of the algorithm. More importantly, participants supported the use of model explanations in all cases, agreeing that even if algorithmic explanations are costly to produce, algorithms should still have explanations. In the case of a hiring algorithm, where there is a clear "human element" to making decisions, participants did not agree that model explanations contributed to information overload, but also did not believe that model explanations were sufficient for resolving sources of confusion. Therefore, while the information presented through model explanations did not overwhelm participants and cloud their decision-making process, there still remained uncertainty in their decisions (see Exhibit 19, C). These results may be driven by the ease of interpreting visual and feature relevance explanations, as it is possible that other explainable AI methods would lead to different outcomes. Additional explainable AI methods would have to be tested to confirm this conjecture.

Scenario I, Qualitative Analysis

Open-ended responses were helpful to observe patterns in participants' attitudes towards model explanations in addition to their suggestions for improving the model. Below are a few observations that add more granularity to the results achieved through this experiment.

Participants given model explanations better recognize the algorithm's error space, while participants in the unexplainable group only notice poor results

When reporting both their surprises and improvements for the algorithm, the explainable participant group correctly identified the only intended "faulty" feature of the algorithm as its image highlighting capability. Most of the explainable participant group cited that the image highlighting "did not accurately assess the candidates that were smiling," or "did not correctly capture the person's smile." In contrast, the unexplainable participant group cited years of experience as the primary reason for surprise in model performance, stating that "the years of experience was a large deciding factor" and "the algorithm recommended they not be hired because of their years of experience." These observations were incorrect, because years of experience were never the "deal-breaking" factor for candidates throughout the simulation. In fact, none of the candidates ever possessed the preferred experience that fit with the company profile, and the algorithm still made a hiring recommendation for Candidate II. The unexplainable participant group conflated the poor performance of the smiling feature to be years of experience, which happened to be the only purely quantitative observation in Scenario I (see Exhibit 20, A).

The explainable participant group also provided a more targeted solution for improving the algorithm, citing procedures such as "the algorithm should get better at picking parts of the images to analyze for specific things like smiling," calling for "more testing of the algorithm, especially computer vision." In contrast, the unexplainable participant group predominantly cited poor results rather than algorithmic processes, such as "I feel as if the third candidate was a better option than the 4th, but the

algorithm gave him the smallest percentage," or "the 3rd candidate should have had a better score in my opinion." (see Exhibit 20, B).

Overall, as observed by Zhang et al. (2020), a better understanding of a model's error space leads to more responsible algorithmic use in the decision-making process, because participants can recognize when they should trust an algorithm's outputs (Zhang et al., 2020, p. 1).

Participants in the explainable group offer substantive evidence for a lack of trust in the model, while the unexplainable group only distrusts the model by its discordance with their own predictions

Participants in the explainable group, having identified the reasons for the model's errors, were precise in their justifications for their eroded trust in its predictions. For example, participants cited the faulty image highlighting as a deal-breaker for their trust in the algorithm, as one participant states, "there were two instances where the AI could not identify the candidate's face and therefore gave them a poor smile rating, which led to the AI to decide that otherwise worthy candidates did not deserve the job. I certainly wouldn't trust this AI to make hiring decisions for the HR department of a company I manage, and I wouldn't take the recommendations into consideration when reviewing applicants." Now compare these observations with the responses from the unexplainable participant group, such as "I could not predict what the algorithm was going to recommend and it made me lose trust in the algorithm's decisions," "the algorithm seemed to often disagree with my thinking," and "sometimes I thought a candidate was perfect but the algorithm said they would not make a good hire" (see Exhibit 20, C). Both participant groups had to make their own judgments based on the information provided, but with a better understanding of the model's error space, the explainable group demonstrated an ability to predict why an algorithm made its predictions, and then cited those reasons as evidence against the validity of a model. If taken through a guided debugging process as conducted in Ribiero et al. (2016), there is reason to believe that the explainable participant group would improve the model's performance (Ribiero et al., 2016, p. 8).

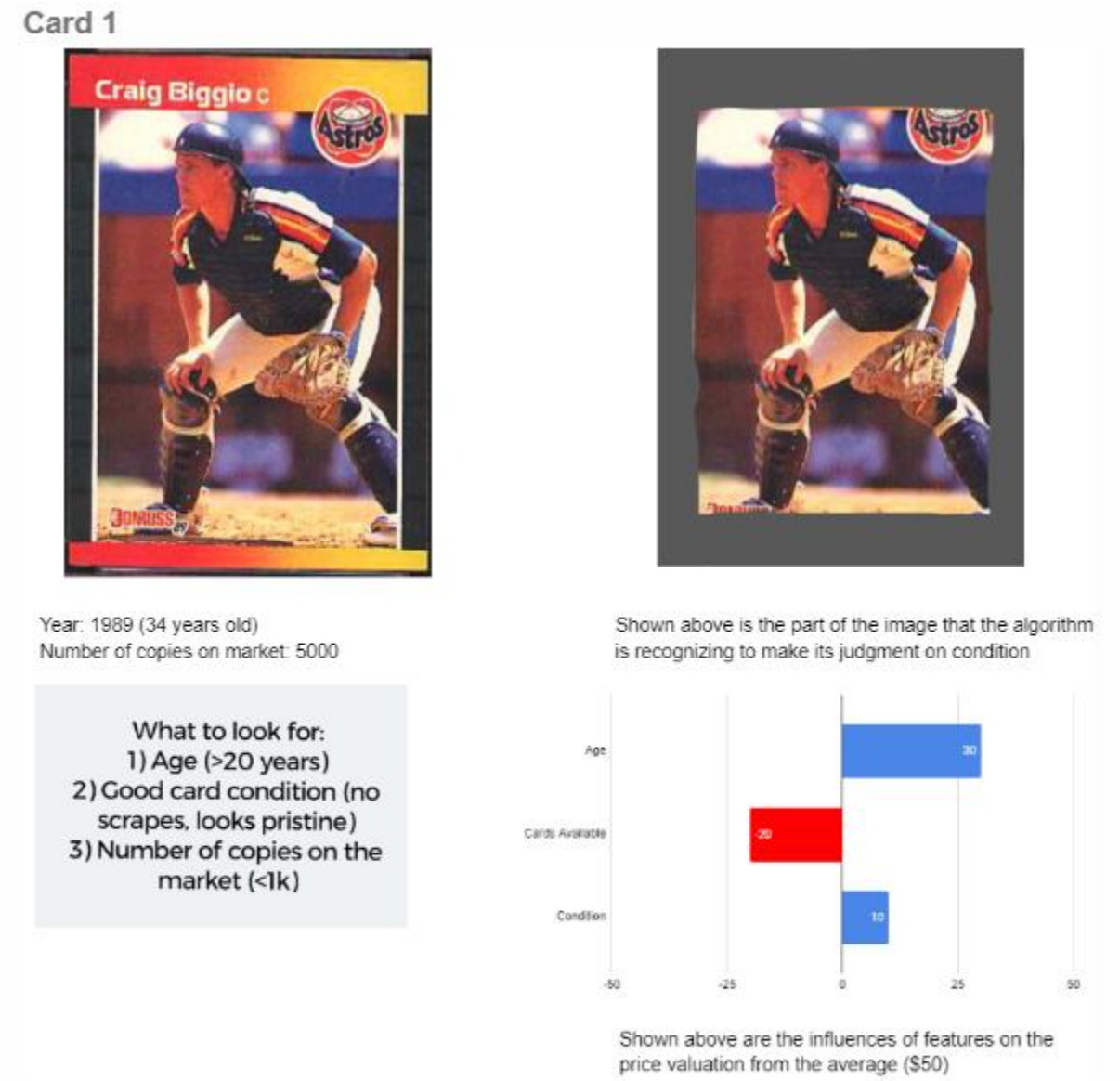
Participants not shown model explanations desire explanations, but participants shown model explanations ask for more information about the explanations

Even though model explanations are intended to provide an explanation for an algorithm's output, which is most useful for error cases, participants in the explainable group asked for more information related to the model explanations, as seen in Liao et al. (2021) (Liao et al., 2021, p. 5). While it seems like the explainable group understood the faults of the algorithm's image highlighting capability, they also expressed dissatisfaction with the limitations of the provided explanations. For example, a few participants stated that "the AI should be able to give more human-like reasoning for how it arrived at its conclusion," "I don't understand how it was matching up synonyms," and "I think more in-depth explanations could be helpful; the more information provided the better" (see Exhibit 20, D). Model explanations were engineered to cater to a diverse group of participants, which left some participants dissatisfied with the perceived incompleteness of the information provided. More studies would have to be done to isolate the effect in which participants reject the adoption of an algorithm's explanations because they believe it did not provide enough information.

The unexplainable participant group wished for more insights into the algorithm's internal calculations, raising questions that are ideally suited to be answered by model explanations. For example, participants request information regarding the "scoring process of each requirement and the weightings of each" and "how the algorithm made its decisions" (see Exhibit 20, D). These requests directly fit with the image highlighting and feature influence bar charts provided to the explainable participant group, which helped participants understand the weights of each feature and how the algorithm detects a candidate's smile.

While the explainable group demands more answers surrounding model explanations, future studies could be conducted to maximize adoption of explainable AI procedures. However, for this study, it is clear that model explanations, while imperfect, still led to increases in performance, adoption, and confidence in a decision-making environment that simulated high levels of risk.

Scenario II, Card I



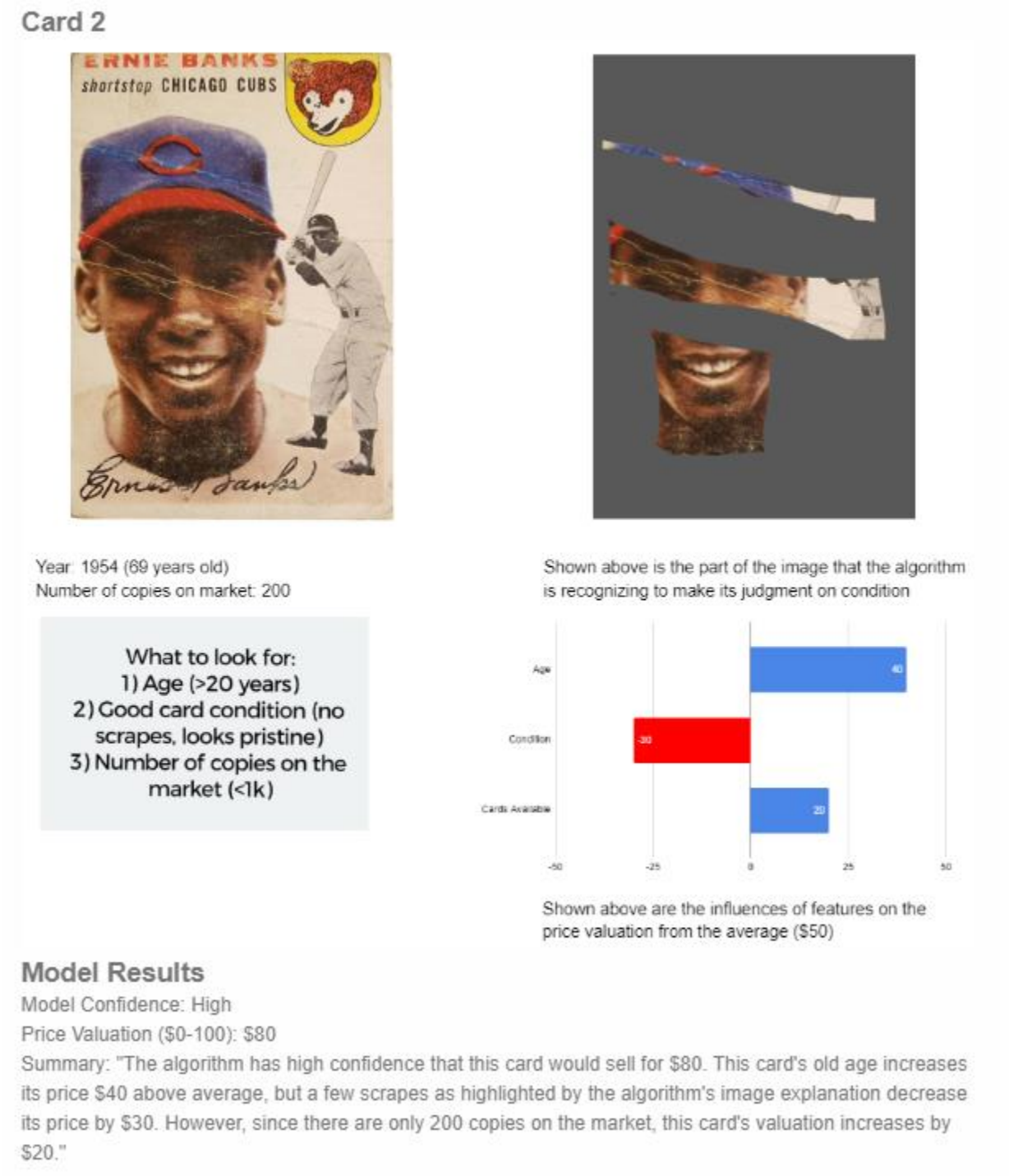
Similar to Candidate I from Scenario I, Card I was meant to be a straightforward evaluation of a baseball card in which people would anchor on the listing price and then make adjustments throughout the survey. Card I is clearly shown to be an old card, as it is well above the 20-year threshold at 34 years old.

However, Card I has numerous copies on the market, which would negatively impact its price. The image highlighting is directly on the picture of the card, which is pristine. Since the age and cards available features are in opposite directions in addition to the fact that this is the first card shown to participants, it was expected that both participant groups would mimic the algorithm's output and sell the card at $70.

A first observation from this case is that 58% of the explainable participants opted to sell the card at the algorithm's recommendation of $70, while a slight majority of the unexplainable participant group were not comfortable selling the card at the algorithm's price. However, while the difference between these proportions was insignificant, the explainable participant group reached a significant majority in their decision to accept the algorithm's recommendation. Explainable participants who did not accept the algorithm's listing price averaged a $57 list price, which is still an above average valuation but significantly lower than $70. Unexplainable group participants who rejected the algorithm's recommendation averaged a $55 list price, which was similar to the explainable group (see Exhibit 21, A). Both groups cited that they balanced their decision with the algorithm at a high rate, at 79% for the explainable group and 80% for the unexplainable group (see Exhibit 21, B). However, confidence scores for the unexplainable group's recommendation were higher than the explainable group, even though they had less information. Confidence scores were high and similar across both groups towards the algorithm's recommendation and ability to generalize to evaluate new cards (see Exhibit 21, C). The feature influence chart was cited as the most helpful explanation method, as it breaks down the effects of each feature on the algorithm's calculations (see Exhibit 21, D).

Participants were likely to anchor on the algorithm's listing price for Card I if they were shown model explanations, which was not replicated by the unexplainable group. A high proportion of participants across both groups used information from the algorithm to support their decision, but more cases need to be studied to determine a difference between groups in their trust of the algorithm's recommendations.

Scenario II, Card II



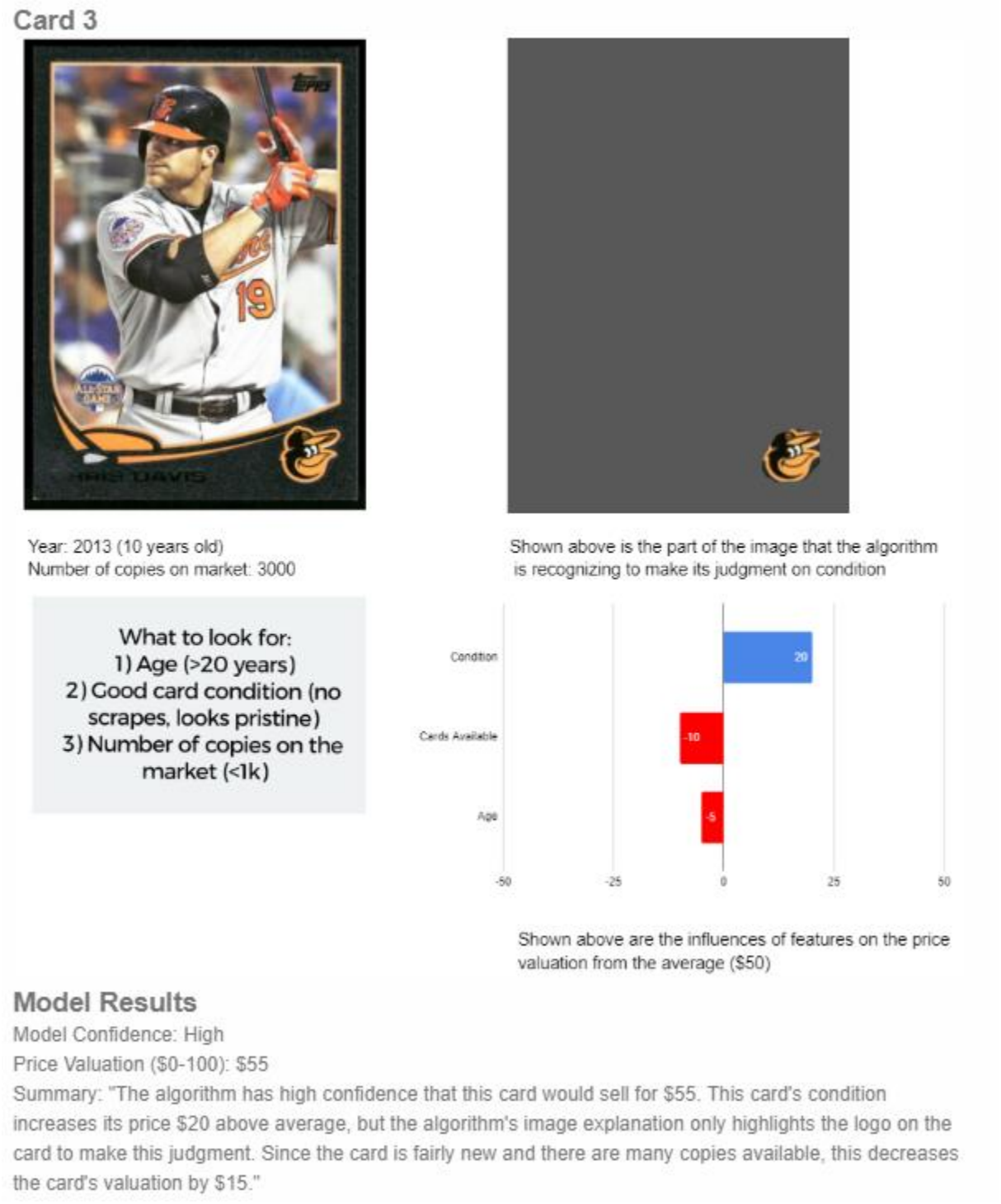
Card II was also supposed to be a fairly straightforward case, with the introduction of scrapes in order to test participants' quantification of a card's condition. For Card II, the image highlighting accurately captures the scrapes on the card, and decreases the algorithm's recommended price for auction,

as shown on the feature influence chart. Conversely, the card's very old age of 69 years paired with its few number of copies on the market at 200 increases its recommended price by the algorithm to $80. While in the hiring algorithm, interpretation of candidate images may offer some confusion, it is clear that Card II is damaged, with scrapes along the player's face. Compared to the previous algorithm's recommendation of $70, it is reasonable to believe that Card II's immense increase in age paired with its rarity would increase its worth relative to Card I, which is the expected behavior from both groups.

Participants in both groups did not support the algorithm's recommendation, with 58% of the explainable group refusing to sell the card at its high price, while 65% of the unexplainable model group also went against the algorithm's pricing recommendation. However, in both cases, the mean value of the participants' listing prices were similar, at $81 for the explainable group and $79 for the unexplainable group (see Exhibit 22, A). Both groups balanced their decisions with the algorithm's recommendation, with 72% balancing the algorithm from the explainable group compared to 65% for the unexplainable group (see Exhibit 22, B). Confidence in both participants' selling prices and the algorithm's decisions remained high for this example, but the confidence in the participants' decisions and the algorithm's performance were stronger within the unexplainable participant group. Compared to Card I's results, only the confidence in the participants' decisions increased while confidence in the algorithm stayed stagnant for both groups (see Exhibit 22, C).

Card II was an attempt to vary the only potential "human" element of this scenario, in that participants may experience confusion when evaluating the condition of baseball cards. However, while it is true that both groups reject the algorithm's recommendation, the price differentials do not offer enough support that participants disagree with the algorithm's price listing. In both explainable and unexplainable groups, results remain similar, in both their decisions and usage rates of the algorithm.

Scenario II, Card III



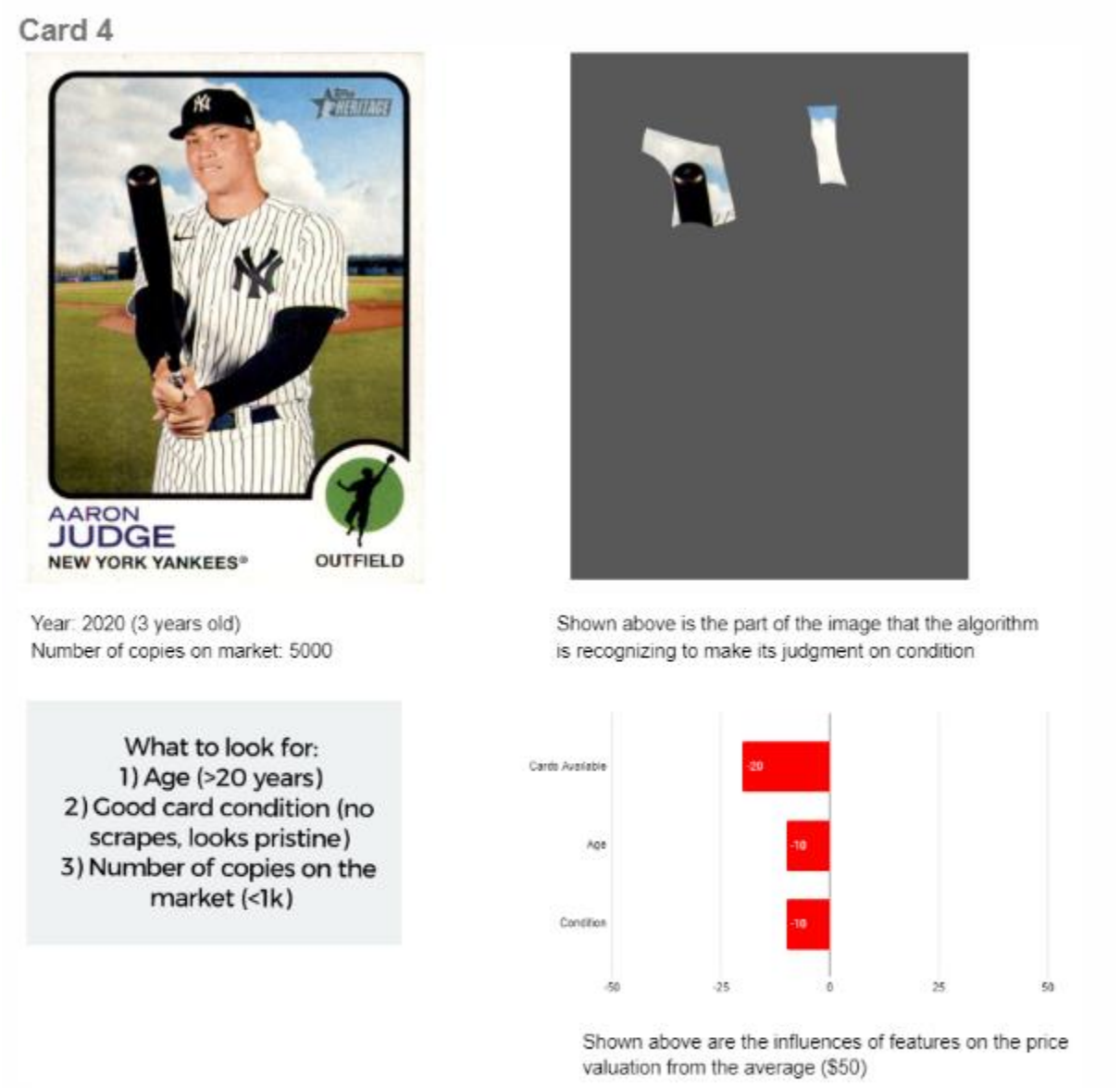
Similar to Candidate III from Scenario I, Card III replicates a poorly performing visual inspection by the algorithm, as the image highlighting reveals that the algorithm failed to scan the image properly on the card. However, the feature influence chart shows that even though an error is made by the algorithm,

it makes an error in the right direction, because it still rates the card as in pristine condition. This observation was expected to lead to distrust in the algorithm from the explainable participants. With a 10-year-old card that has 3000 copies on the market, Card III was intended to be accurately priced at $55. As such, this example would reveal if participants would still trust an algorithm's recommendation, even if the algorithm's "error" was not numerically meaningful.

Across both explainable and unexplainable participant groups, the majority voted against the algorithm's recommendation, with 84% of the explainable group lowering the price to $36, and 82% of the unexplainable group lowering the price to $35, which yields similar results (see Exhibit 23, A). Algorithmic use is also similar across both groups, as 48% of the explainable group balanced the algorithm's output with their decision, while this proportion was 47% for the unexplainable group (see Exhibit 23, B). Both groups rate a high confidence level in their own recommendation but have lower confidence in both the algorithm's prediction for Card III and its ability to generalize. A surprising observation is that confidence ratings of the algorithm are the same across both groups, even though the unexplainable group did not have access to the image highlighting errors (see Exhibit 23, C).

While the algorithm displayed a mistake in Card III, it did not contribute to an error in its final recommendation, which may have kept trust in the algorithm from the explainable participant group. Consistent with the findings from Ribiero (2018), Card III can be reasoned as a case in which the algorithm was "not acting in an unreasonable manner," since the card was still analyzed as pristine by the algorithm (Ribiero et al., 2016, p. 5).

Scenario II, Card IV



Card IV is another example of an algorithmic mistake in its image recognition capabilities as the image highlighting focuses on portions of the card that are not damaged. Therefore, the displayed

negative effect on the card's recommended price by the algorithm is inaccurate, because there are no scrapes on the card. Having seen the mistake of the image recognition form Card III, the explainable group was expected to be primed to distrust the algorithm's recommendation and change their price based on the faulty image highlighting. Compared to Card III, Card IV is a newer card with more cards on the market, which would decrease its value relative to Card III. As such, it is expected that Card IV would be priced at a lower price than Card III's $55, but higher than the $10 list price.

The explainable participant group rejected the algorithm's recommendation at a higher rate compared to the unexplainable group, as 56% of explainable participants chose to set their own price, compared to only 32% of the unexplainable population group. For both groups, the direction of the price change was in the correct direction, and less than Card III, with similar means of $20 and $17 for the explainable and unexplainable groups, respectively (see Exhibit 24, A). While the majority of both groups balanced their decision with the algorithm's output, a larger portion of the unexplainable group defaulted to the algorithm's output, while the explainable group made the decision on their own at a higher rate (see Exhibit 24, B). Confidence in participants' decisions differed across all categories, as the unexplainable participant group expressed higher confidence in both their own recommendations as well as the algorithm's recommendation and generalization ability (see Exhibit 24, C).

Card IV is a case in which the explainable group appropriately identifies the algorithm's mistake, while conversely, the majority of the unexplainable group simply accepts the algorithm's recommendation. However, the proposed price changes by both groups remained similar, which means that the rejection of Card IV's recommendation by the explainable group was not influenced in a particular direction.

Scenario II, End of Survey Results

Addressing the demographics in this survey, the knowledge of AI was below average for the explainable group at a self-rating of 40; however, the unexplainable participant group had a higher self-rating of 47. While this difference was significant, these measures were comparable to Scenario I. In

Scenario II, domain knowledge of baseball would be important as the players on each card may have tipped pricing decisions in a certain direction, but this self-reported rating was low and similar between groups, at 35 and 36 for both explainable and unexplainable groups, respectively (see Exhibit 25, A). Both groups reported a moderate degree in which the algorithm influenced their decision making, but the explainable participant group was more influenced by the algorithm, at a rating of 61 compared to 56 for the unexplainable group (see Exhibit 25, B). It is worth noting that both groups agreed that they had enough information to make decisions on pricing baseball cards with confidence (see Exhibit 25, C).

For the explainable group, similar to Scenario I, there was agreement that model explanations helped them better understand the capabilities of the algorithm. In fact, participants in the explainable group appreciated the model explanations to the degree in which they agreed that algorithms should always have explanations, even if they are costly to produce. While in a quantitative exercise, model explanations are expected to contribute to information overload as suggested in Zhang et al. (2020), but participants did not believe that this was the case (Zhang et al., 2020, p. 9). In fact, a key difference compared to Scenario I is that the explainable group in Scenario II believed that model explanations were sufficient for resolving any sources of confusion during the simulation (see Exhibit 25, C).

Scenario II, Qualitative Analysis

As previously mentioned, open-ended responses were helpful to observe patterns in attitudes towards model explanations and understand suggestions for improving the model. Below are a few observations that offer support for the above claims in this paper.

Participants given model explanations better recognize the algorithm's error space, while participants in the unexplainable group only notice poor results

Similar to Scenario I, participants in the explainable group were able to identify the shortcomings of the algorithm's image highlighting capabilities, which is consistent with the findings of Zhang et al. (2020) regarding participants' learning of an algorithm's error space (Zhang et al., 2020, p. 1). For

example, participants in the explainable group stated that “the image highlighting wasn’t always the most accurate,” “the image highlighting prompted deductions that weren’t necessarily accurate,” and “it spotted scratches that didn’t seem to exist.” In contrast, the unexplainable group believed that the algorithm was conservative, even though the algorithm was pricing cards at both the high end ($80 for Card II) and the low end ($10 for Card IV) of the list price spectrum. A few participants claimed that “the algorithm made fairly modest valuations” and “seemed to make conservative results that stayed closer to the average price” (see Exhibit 26, A).

Similar to the observations of Ribiero et al. (2016), there is reason to believe that the explainable participant group would improve the model’s performance through a guided debugging process (Ribiero et al., 2016, p. 8). Participants in the explainable group also suggested reasonable improvements to the image highlighting process that could lead to a suitable debugging process, while the unexplainable participant group just cited poor results in the algorithm’s output, which also replicates results from Scenario I. For example, explainable participants noticed that the image highlighting focused on the wrong parts of the card and was rather inaccurate in its interpretation of card condition. Participants shown model explanations stated that image highlighting “could make pricing lower than market value” and had a “focus on specific aspects like logos.” In contrast, the participants from the unexplainable model group simply reported the alignment of the algorithm’s results with their own view, as one participant stated that “I thought the algorithm was more accurate for the first 2 cards, but less accurate for the last two” (see Exhibit 26, B). Therefore, while the explainable participant group offered informed responses for improving the algorithm, the unexplainable participant group was not able to identify the source of the algorithm’s errors.

Participants are slightly confused by model explanations

In a more quantitative exercise compared to Scenario I, participants shown model explanations experienced some confusion around the algorithm’s process for explanations. Participants seemed to misinterpret the image highlighting as only scanning a few portions of the card instead of highlighting the

most important pixels in its complete scan of the card's features. For example, participants claimed that they "didn't completely understand why the algorithm couldn't read the entire image," and that it was "making decisions on partial images." One participant seemed to struggle with the feature influence bar chart, stating that "the bar chart was a bit hard to read at first so it took me longer to understand/comprehend the data being presented" (see Exhibit 26, C). While there may be some reason to believe that these struggles reflect an information overload in Scenario II, more studies need to be done to understand the threshold of when information overload is reached, and if such factors are dependent on the type of model explanation provided.

Participants not shown model explanations desire explanations, but participants shown model explanations ask for more information about the explanations

As mentioned by Miller (2018), Liao et al. (2021), and also observed in Scenario I, model explanations do not yet offer complete and comprehensive explanations (Liao et al., 2021, p. 5; Miller, 2018, p. 6). Therefore, the image highlighting and feature influence bar charts, while informative, did not satisfy some participants in Scenario II. For example, participants stated that they "want reasoning as to why specific amounts of deductions were issued," and wanted "more explanation for why certain parts of the image were highlighted versus others." Participants also wanted explanations by counterfactual, putting "comparable valuations, previously sold cards," and "average cards to compare" on their wish lists for improving model explanations (see Exhibit 26, D).

Participants in the unexplainable group also desired more information, but they primarily cited information that would be contained in a model explanation. For example, participants asked, "how much did each piece of criteria factor into the price? What was each categories' worth?" Additionally, participants were curious to know "how the algorithm evaluated the condition of the card." Overall, participants in the unexplainable group wanted to "get a better understanding of how the algorithm arrived at each price," which was the primary purpose of providing model explanations in Scenario II (see Exhibit 26, D).

VI. Conclusion

In cases involving high-risk consequences from algorithmic mistakes, it is clear that explainable AI techniques positively impact performance for AI non-experts. Participants shown model explanations have a better understanding of an algorithm's error space, which is calibrated quickly over only a few examples. More importantly, when presented with underlying algorithmic errors, participants display more responsible algorithmic use, choosing to diverge from an algorithm's output more often in error cases. This behavior creates a more informative attitude towards the potential danger of an algorithm's recommendations, which was reflected by sharp and continued drops in confidence in the hiring algorithm example. Therefore, the information presented through model explanations, while static, served its intended purpose as a safeguard against poor outcomes in "human" problem spaces of high social risk.

However, in more quantitative, straightforward cases, explainable AI procedures offer little benefit over a black box system. While participants were able to understand the error space of the algorithm and reject its recommendations when necessary, they did not override the algorithm's recommendations in a specific direction or a numerically meaningful way. Both groups in Scenario II offered similar prices across all examples, which illustrates negligent effects from algorithmic aversion. This similarity means that in practice, it may not be worth the extra cost to produce model explanations in quantitative, straightforward cases, especially when considering the added risks of information overload and confusion for decision makers.

Therefore, it is reasonable to conclude that in use cases involving high risk, explainable AI should be the primary tool used for decision support. Explainable AI systems can help safeguard against harmful failure cases and accelerate a decision maker's learning of an algorithm's capabilities. When discussing the implications of AI systems "in the wild," responsible deployment requires fast and iterative learning, in which explainable AI algorithms can offer both faster and more informative insights about an algorithm's performance.

A few limitations of this study involve a lack of testing for the most effective explainable AI techniques, coverage for the disagreement problem, and the potential of fair washing. Future studies can

be conducted to enhance the user experience with explainable AI techniques, aligning the field's social task of explanation with AI's progress towards conversational, language-based interfaces. Additionally, more work towards understanding behavior of decision makers when explanations disagree should be studied to resolve sources of conflict not only with a model's recommendations, but the explainable AI's interpretation of those recommendations. Third, studies can be done to allow decision makers to debug an explainable algorithm, which can inform studies on long-term performance as well as the risks for "hacking" an algorithm to serve malicious ends.

References

Anderson, J., Rainie, L., & Cohn, S. (2022). *Artificial Intelligence and the Future of Humans*. Pew Research Center: Internet, Science & Tech. Retrieved April 17, 2023, from https://www.pewresearch.org/internet/2018/12/10/artificial-intelligence-and-the-future-of-humans/

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Explainable AI. IBM. (n.d.). Retrieved October 27, 2022, from https://www.ibm.com/watson/explainable-ai

Harwell, D. (2019). A face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post*. Retrieved February 20, 2023, from https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*. Retrieved February 19, 2023, from https://link.springer.com/article/10.1007/s12525-021-00475-2

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, Z. S., & Lakkaraju, H. (2022). The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *ArXiv*, 1–46.

Liao, Q. V., Gruen, D., & Miller, S. (2021). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *ArXiv*. Retrieved February 20, 2023, from https://arxiv.org/pdf/2001.02478.pdf

McKinsey & Company. (2022). The state of AI in 2022-and a half decade in Review. *McKinsey & Company*. Retrieved February 19, 2023, from https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#/

Miller, T. (2018). Explanation in artificial intelligence: Insights from the Social Sciences. *ArXiv*, 1–66.

Mittal, N., Saif, I., & Ammanath, B. (2022). Fueling the AI transformation: Four key actions powering widespread value from AI, right now. *Deloitte United States*. Retrieved February 19, 2023, from https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-ai-2022.html

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. https://doi.org/10.18653/v1/n16-3020

Sarker, I. H. (2021). Deep learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *Springer Nature Singapore*. Retrieved February 19, 2023, from https://link.springer.com/article/10.1007/s42979-021-00815-1

Vince. (2022). How to price sports cards. *Sports Cards Rock*. Retrieved February 20, 2023, from https://sportscardsrock.com/how-to-price-sports-cards/

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372852

Appendix

Exhibit 1: Scenario I, Hiring Algorithm Task Introduction

"You are working as an intern for a human resources group, and it is your responsibility to vet candidates for the position of "manager" and make recommendations to your hiring manager. You will be given a briefing card on what the company is looking for in potential candidates, and the same briefing will be used to train an AI algorithm. While doing your job, you are provided the assistance of this AI algorithm. The algorithm shows great promise for vetting candidates, but just like you, it has just been trained and is still being evaluated.

Your task is to view details of 4 candidates and make a hiring recommendation to your line manager for each. You are free to use the algorithm and its information provided as an aid for your decision making, but you are not required to do so. It is imperative that you make the best recommendations to your line manager, or **you risk losing your chance at a full time position with the company after the internship**."

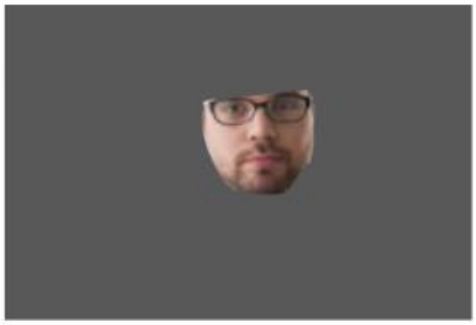
Exhibit 2: Scenario I, Hiring Algorithm Task Briefing

"This company is looking for candidates who fit the following characteristics

- Candidates who are smiling
- Candidates who mention keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
- Candidates who have 5 or more years of experience"

Exhibit 3: Scenario I, Hiring Algorithm Task with Explanations

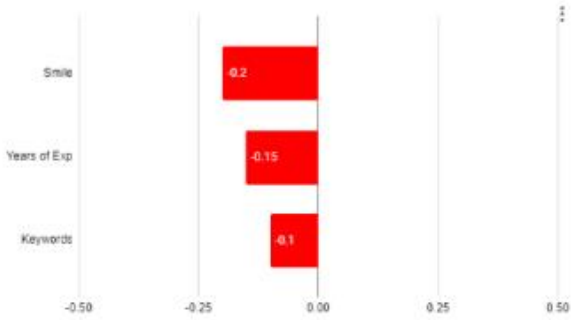
Candidate 1



Keywords: comfortable, rigid, even-keeled, follower
Years of experience: 1

Shown above is the part of the image that the algorithm is recognizing to make its judgment on smile

What to look for:
1) A smiling candidate
2) Keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
3) 5 or more years of experience



Shown above are the influences of features on candidate score from the average (50%)

Model Results

Model Confidence: High
Candidate Score (0-100%): 5%
Recommendation: Do not hire
Summary: "The algorithm has high confidence that this person would not make a good hire. A lack of a smile, as highlighted by the algorithm's image explanation, contributes to a 20% decrease in the probability that this person would make a good hire, followed by few years of experience and poor keywords, at 15% and 10%, respectively."

Exhibit 4: Scenario I, Hiring Algorithm Task with No Explanations

Candidate 1

Keywords: comfortable, rigid, even-keeled, follower
Years of experience: 1

What to look for:
1) A smiling candidate
2) Keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
3) 5 or more years of experience

Model Results

Model Confidence: High
Candidate Score (0-100%): 5%
Recommendation: Do not hire
Summary: "The algorithm has high confidence that this person would not make a good hire."

Exhibit 5: Scenario I, Hiring Algorithm Task, Questions after each Hiring Decision

Participant group provided with model explanations:

1) Would you recommend hiring this candidate?
    a) Yes
    b) No
2) How did you arrive at your recommendation?

   a) I made this decision on my own
   b) I balanced my decision with information from the algorithm
   c) I defaulted to the algorithm's recommendation
3) Of the model explanations provided, which one helped you more in making your recommendation?
   a) Image highlighting
   b) Feature relevance bar chart
4) Please rate your confidence for the following (Scale 0-100, 0 = not confident, 100 = very confident):
   a) Please rate your confidence that you made the correct hiring recommendation
   b) Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example
   c) Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general

Participant group provided with no model explanations:

1) Would you recommend hiring this candidate?
   a) Yes
   b) No
2) How did you arrive at your recommendation?
   a) I made this decision on my own
   b) I balanced my decision with information from the algorithm
   c) I defaulted to the algorithm's recommendation
3) Please rate your confidence for the following (Scale 0-100, 0 = not confident, 100 = very confident):
   a) Please rate your confidence that you made the correct hiring recommendation

b) Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example

c) Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general

Exhibit 6: Scenario I, Hiring Algorithm Task, Reflection Questions at end of Survey

Participant group provided with model explanations:

1) Please answer the following:
    a) Please rate your knowledge of artificial intelligence (Scale 0-100, 0 = little knowledge, 100 = expert knowledge)
    b) Please rate the degree in which model explanations influenced your decision making (Scale 0-100, 0 = not at all, 100 = a lot)
2) Please rate your agreement or disagreement with the following statements about model explanations (image highlighting, feature influence) (Likert scale for agreement):
    a) I had enough information to make hiring recommendations with confidence.
    b) Model explanations helped me better understand the strengths and weaknesses of the algorithm.
    c) Algorithms should always have explanations to help decision makers, even if they are costly to produce.
    d) Model explanations offered too much information and confused me in the decision-making process.
    e) The model explanations provided were sufficient for resolving sources of confusion.
3) Was there anything you noticed through the model explanations (image highlighting, feature influence) that surprised you? Why? (Open-ended response)
4) Is there anything you would like to see improve from the model explanations? (Open-ended response)

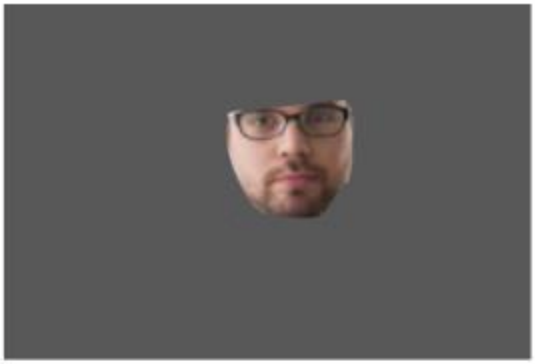
Participant group provided with no model explanations:

1) Please answer the following:
    a) Please rate your knowledge of artificial intelligence (Scale 0-100, 0 = little knowledge, 100 = expert knowledge)
    b) Please rate the degree in which the algorithm influenced your decision making (Scale 0-100, 0 = not at all, 100 = a lot)
2) Please rate your agreement or disagreement with the following statement (Likert scale for agreement):
    a) I had enough information to make hiring recommendations with confidence.
3) Was there anything you noticed about the results of the algorithm that surprised you? Why? (Open-ended response)
4) During the study, was there anything you would have liked to know about the algorithm? If so, please explain. (Open-ended response)

Exhibit 7: Scenario I, Hiring Algorithm Task, Candidates

Candidate I, with model explanations:

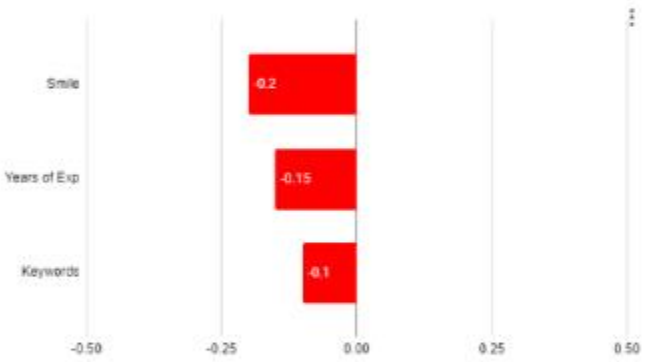
Candidate 1



Keywords: comfortable, rigid, even-keeled, follower
Years of experience: 1

Shown above is the part of the image that the algorithm is recognizing to make its judgment on smile

What to look for:
1) A smiling candidate
2) Keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
3) 5 or more years of experience

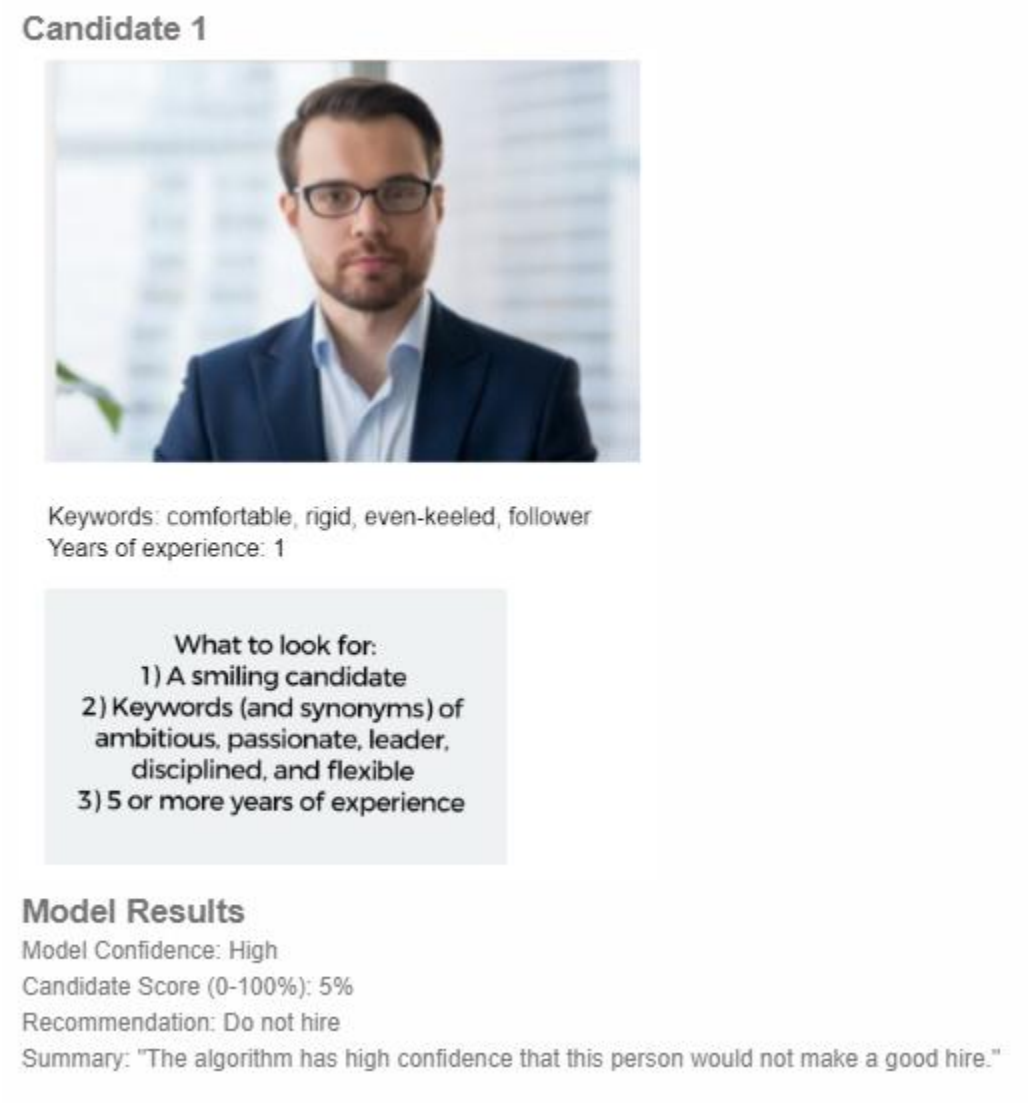


Shown above are the influences of features on candidate score from the average (50%)

Model Results

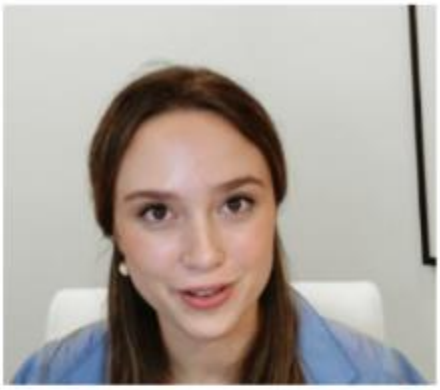
Model Confidence: High
Candidate Score (0-100%): 5%
Recommendation: Do not hire
Summary: "The algorithm has high confidence that this person would not make a good hire. A lack of a smile, as highlighted by the algorithm's image explanation, contributes to a 20% decrease in the probability that this person would make a good hire, followed by few years of experience and poor keywords, at 15% and 10%, respectively."

Candidate I, with no model explanations:



Candidate II, with model explanations:

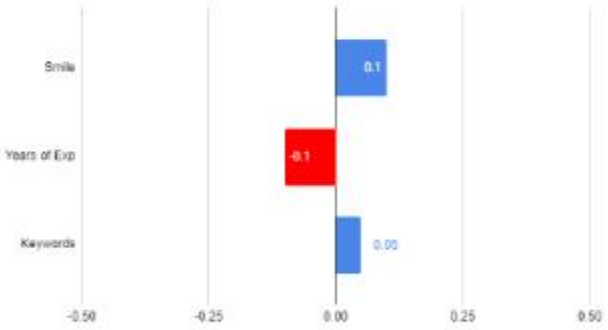
Candidate 2



Keywords: reserved, gracious, disciplined, hard-working
Years of experience: 3

Shown above is the part of the image that the algorithm is recognizing to make its judgment on smile

What to look for:
1) A smiling candidate
2) Keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
3) 5 or more years of experience



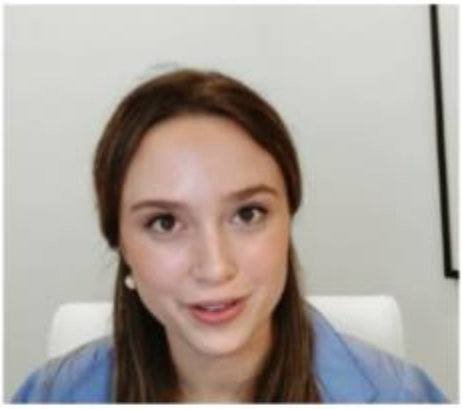
Shown above are the influences of features on candidate score from the average (50%)

Model Results

Model Confidence: High
Candidate Score (0-100%): 55%
Recommendation: Hire
Summary: "The algorithm has high confidence that this person would make a good hire. While the years of experience cause a 10% decrease in the probability of making a good hire, the smile (as highlighted by the algorithm's image explanation) cancels this effect. Keywords like 'disciplined' are what the company is looking for, boosting the probability that this person would make a good hire by 5%."

Candidate II, with no model explanations:

Candidate 2



Keywords: reserved, gracious, disciplined, hard-working
Years of experience: 3

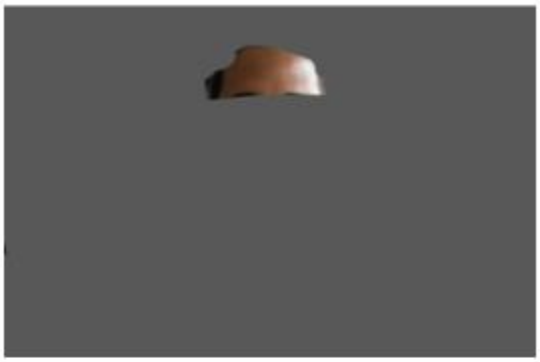
What to look for:
1) A smiling candidate
2) Keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
3) 5 or more years of experience

Model Results

Model Confidence: High
Candidate Score (0-100%): 55%
Recommendation: Hire
Summary: "The algorithm has high confidence that this person would make a good hire."

Candidate III, with model explanations:

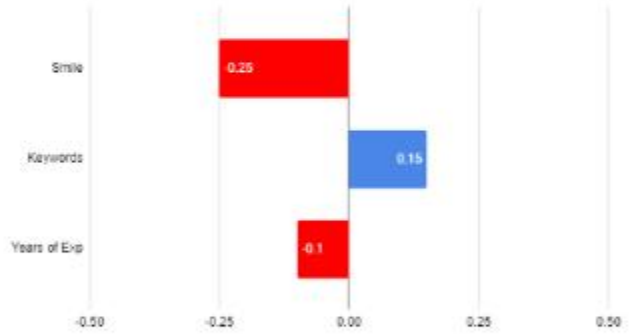
Candidate 3



Keywords: upbeat, growth mindset, leader, rule follower
Years of experience: 3

Shown above is the part of the image that the algorithm is recognizing to make its judgment on smile

What to look for:
1) A smiling candidate
2) Keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
3) 5 or more years of experience

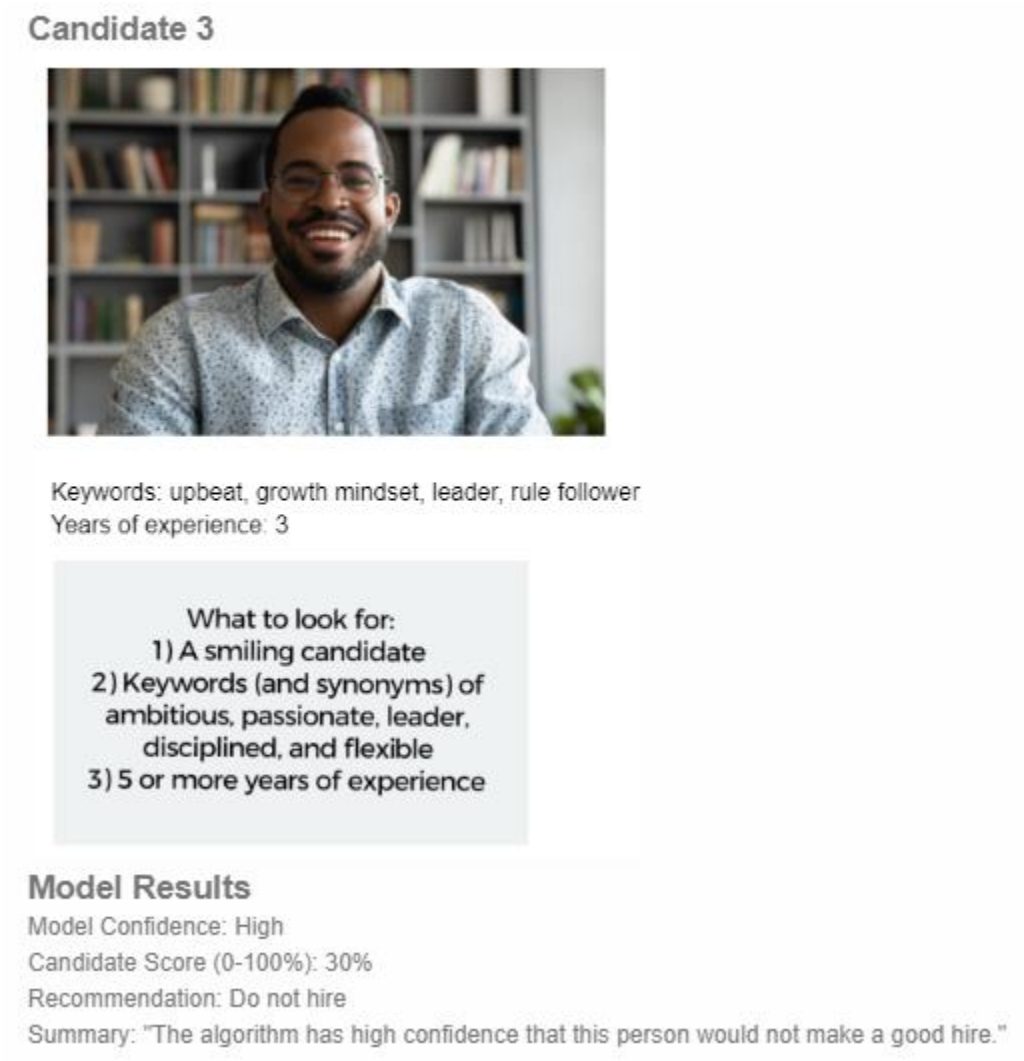


Shown above are the influences of features on candidate score from the average (50%)

Model Results

Model Confidence: High
Candidate Score (0-100%): 30%
Recommendation: Do not hire
Summary: "The algorithm has high confidence that this person would not make a good hire. As shown by the image explanation, the algorithm's failure to capture the smile of the candidate leads to a 25% decrease in probability that they would be a good hire. The candidate's keywords are strong, leading to a 15% boost in probability of making a good hire, but their lack of experience penalizes most of that effect."

Candidate III, with no model explanations:



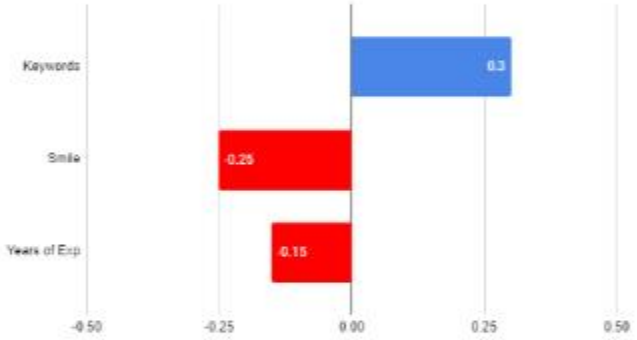
Candidate IV, with model explanations:

Candidate 4

Keywords: ambitious, disciplined, leader, positive
Years of experience: 1

Shown above is the part of the image that the algorithm is recognizing to make its judgment on smile

What to look for:
1) A smiling candidate
2) Keywords (and synonyms) of ambitious, passionate, leader, disciplined, and flexible
3) 5 or more years of experience



Shown above are the influences of features on candidate score from the average (50%)

Model Results

Model Confidence: High
Candidate Score (0-100%): 40%
Recommendation: Do not hire
Summary: "The algorithm has high confidence that this person would not make a good hire, as their poor smile decreases the probability that they would make a good hire by 25%. However, the algorithm, as shown by the image explanation, does not capture the smile, but rather the background of the image. Additionally, only one year of experience decreases their probability of being a good hire by 15%, but strong keywords boost this probability by 30%."

Candidate IV, with no model explanations:

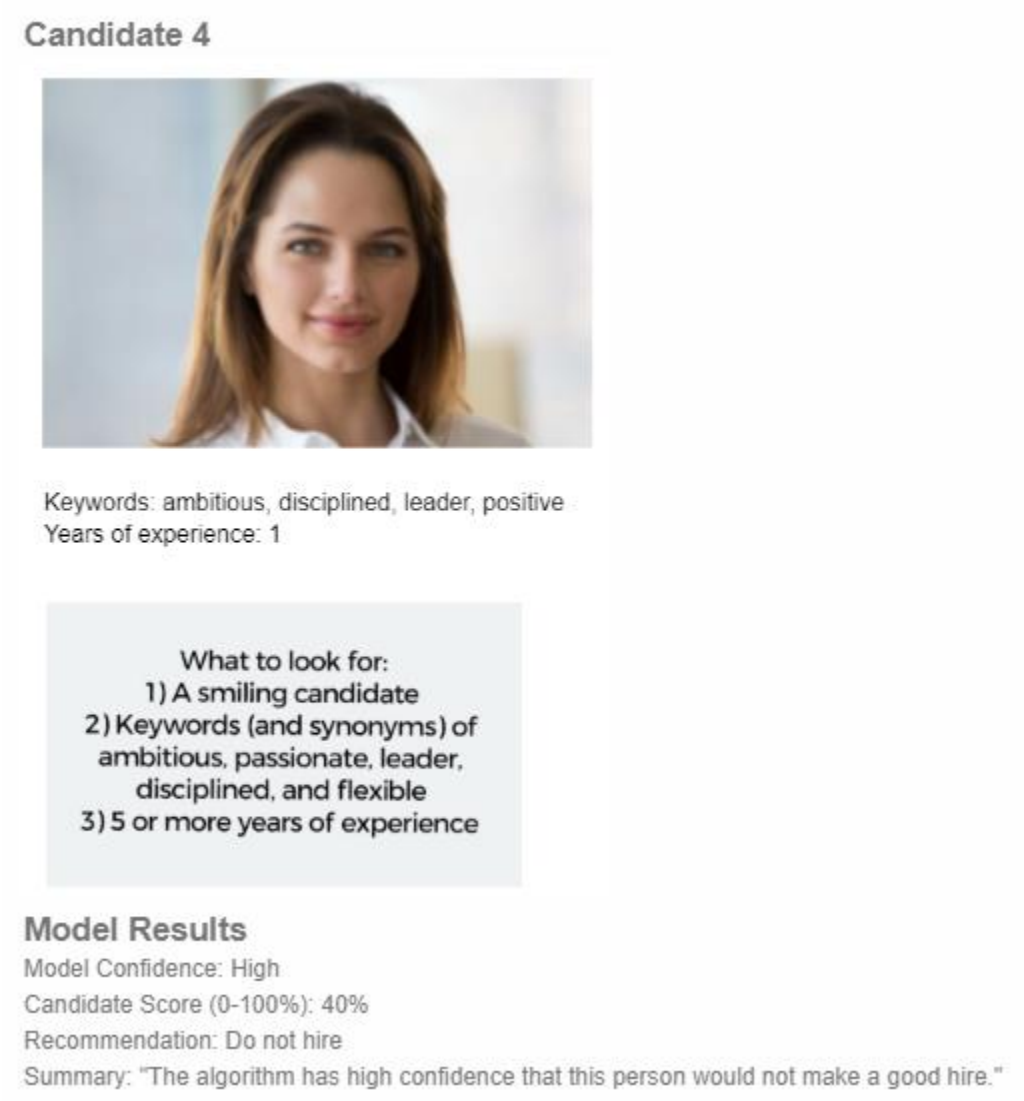


Exhibit 8: Scenario II, Baseball Card Pricing Algorithm Task Briefing

"A valid baseball card valuation ($0-100, $50 is average) considers the following information

- Age of the card (>20 years is beneficial)
- Condition of the card (no scrapes, pristine look is beneficial)
- Number of copies available on the market (<1000 cards is beneficial)"

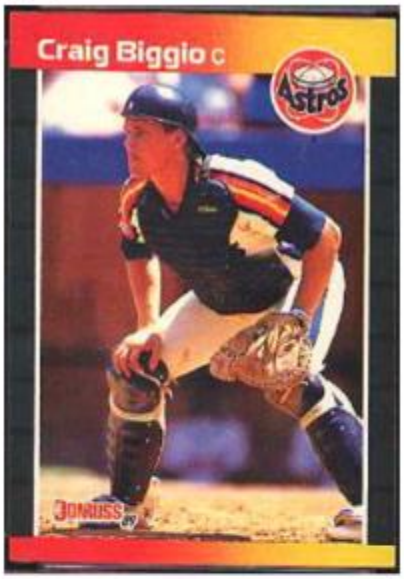
***Exhibit 9:* Scenario II, Baseball Card Pricing Algorithm Task Introduction**

"You are working as an intern for a baseball card show, and it is your responsibility to set selling prices ($0-100) for various baseball cards. You will be given a briefing on how to estimate the price of a card, and the same briefing will be used to train an AI algorithm. While doing your job, you are provided the assistance of this AI algorithm. The algorithm shows great promise for generating effective market prices, but just like you, it has just been trained and is still being evaluated.

Your task is to make price recommendations ($0-100, average is $50) to your line manager for 4 baseball cards. You are free to use the algorithm and its information provided as an aid for your decision making, but you are not required to do so. You want to make appropriate recommendations to your line manager because **you would like them to think highly of you**."

Exhibit 10: ***Scenario II, Baseball Card Pricing Algorithm Task with Explanations***

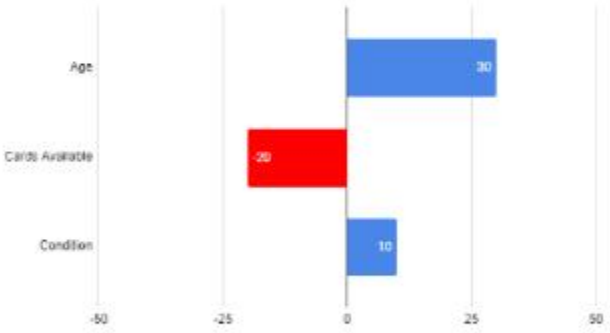
Card 1



Year: 1989 (34 years old)
Number of copies on market: 5000

Shown above is the part of the image that the algorithm is recognizing to make its judgment on condition

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)



Shown above are the influences of features on the price valuation from the average ($50)

Model Results

Model Confidence: High
Price Valuation ($0-100): $70
Summary: "The algorithm has high confidence that this card would sell for $70. The old age of the card boosts its price $30 above average and its clear image as identified by the algorithm's image explanation increases its price by $10. However, the large number of copies on the market decreases its valuation by $20."

Exhibit 11: ***Scenario II, Baseball Card Pricing Algorithm Task with No Explanations***

Card 1

Year: 1989 (34 years old)
Number of copies on market: 5000

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)

Model Results

Model Confidence: High
Price Valuation ($0-100): $70
Summary: "The algorithm has high confidence that this card would sell for $70"

Exhibit 12: Scenario II, Baseball Card Pricing Algorithm Task, Questions after each Pricing Decision

Participant group provided with model explanations:

1) Would you recommend listing this card for $X? (the amount recommended by the algorithm)
    a) Yes

   b) No

2) If you said “No” above, please set your card price below: (Slider, $0-100)
3) How did you arrive at your recommendation?
   a) I made this decision on my own
   b) I balanced my decision with information from the algorithm
   c) I defaulted to the algorithm’s recommendation
4) Of the model explanations provided, which one helped you more in making your recommendation?
   a) Image highlighting
   b) Feature relevance bar chart
5) Please rate your confidence for the following (Scale 0-100, 0 = not confident, 100 = very confident):
   a) Please rate your confidence that you made an appropriate valuation
   b) Please rate your confidence in the algorithm’s ability to make an appropriate valuation for this example
   c) Please rate your confidence in the algorithm’s ability to make appropriate valuations in general

Participant group provided with no model explanations:

1) Would you recommend listing this card for $X? (the amount recommended by the algorithm)
   a) Yes
   b) No
2) If you said “No” above, please set your card price below: (Slider, $0-100)
3) How did you arrive at your recommendation?
   a) I made this decision on my own
   b) I balanced my decision with information from the algorithm

c) I defaulted to the algorithm's recommendation

4) Please rate your confidence for the following (Scale 0-100, 0 = not confident, 100 = very confident):
    a) Please rate your confidence that you made an appropriate valuation
    b) Please rate your confidence in the algorithm's ability to make an appropriate valuation for this example
    c) Please rate your confidence in the algorithm's ability to make appropriate valuations in general

Exhibit 13: Scenario II, Baseball Card Pricing Algorithm Task, Reflection Questions at end of Survey

Participant group provided with model explanations:

1) Please answer the following:
    a) Please rate your knowledge of artificial intelligence (Scale 0-100, 0 = little knowledge, 100 = expert knowledge)
    b) Please rate your knowledge of baseball (Scale 0-100, 0 = little knowledge, 100 = expert knowledge)
    c) Please rate the degree in which model explanations influenced your decision making (Scale 0-100, 0 = not at all, 100 = a lot)
2) Please rate your agreement or disagreement with the following statements about model explanations (image highlighting, feature influence) (Likert scale for agreement):
    a) I had enough information to make card valuations with confidence.
    b) Model explanations helped me better understand the strengths and weaknesses of the algorithm.
    c) Algorithms should always have explanations to help decision makers, even if they are costly to produce.

    d) Model explanations offered too much information and confused me in the decision-making process.
    e) The model explanations provided were sufficient for resolving sources of confusion.
3) Was there anything you noticed through the model explanations (image highlighting, feature influence) that surprised you? Why? (Open-ended response)
4) Is there anything you would like to see improve from the model explanations? (Open-ended response)

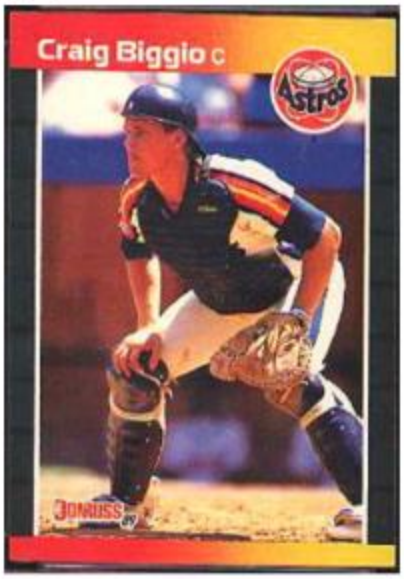
Participant group provided with no model explanations:

1) Please answer the following:
    a) Please rate your knowledge of artificial intelligence (Scale 0-100, 0 = little knowledge, 100 = expert knowledge)
    b) Please rate your knowledge of baseball (Scale 0-100, 0 = little knowledge, 100 = expert knowledge)
    c) Please rate the degree in which the algorithm influenced your decision making (Scale 0-100, 0 = not at all, 100 = a lot)
2) Please rate your agreement or disagreement with the following statement (Likert scale for agreement):
    a) I had enough information to make card valuations with confidence.
3) Was there anything you noticed about the results of the algorithm that surprised you? Why? (Open-ended response)
4) During the study, was there anything you would have liked to know about the algorithm? If so, please explain. (Open-ended response)

Exhibit 14: Scenario II, Baseball Card Pricing Algorithm Task, Cards

Card I, with model explanations:

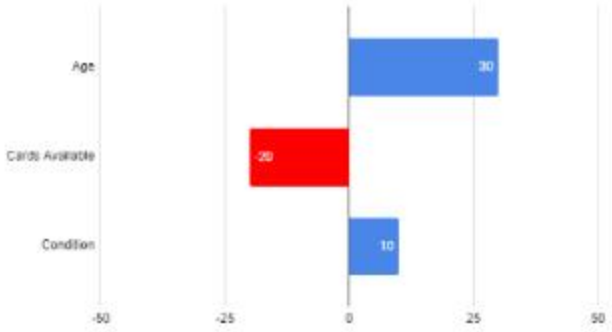
Card 1



Year: 1989 (34 years old)
Number of copies on market: 5000

Shown above is the part of the image that the algorithm is recognizing to make its judgment on condition

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)



Shown above are the influences of features on the price valuation from the average ($50)

Model Results

Model Confidence: High
Price Valuation ($0-100): $70
Summary: "The algorithm has high confidence that this card would sell for $70. The old age of the card boosts its price $30 above average and its clear image as identified by the algorithm's image explanation increases its price by $10. However, the large number of copies on the market decreases its valuation by $20."

Card I, with no model explanations:

Card 1

Year: 1989 (34 years old)
Number of copies on market: 5000

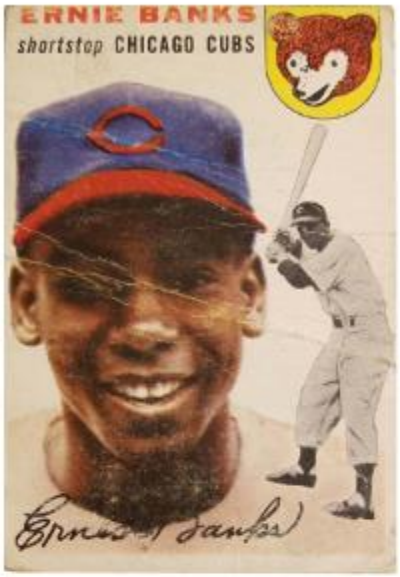
What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)

Model Results

Model Confidence: High
Price Valuation ($0-100): $70
Summary: "The algorithm has high confidence that this card would sell for $70"

Card II, with model explanations:

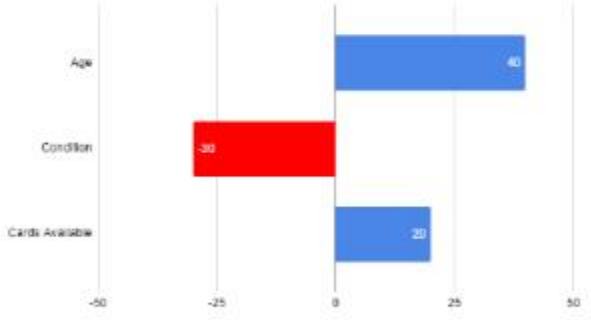
Card 2



Year: 1954 (69 years old)
Number of copies on market: 200

Shown above is the part of the image that the algorithm is recognizing to make its judgment on condition

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)

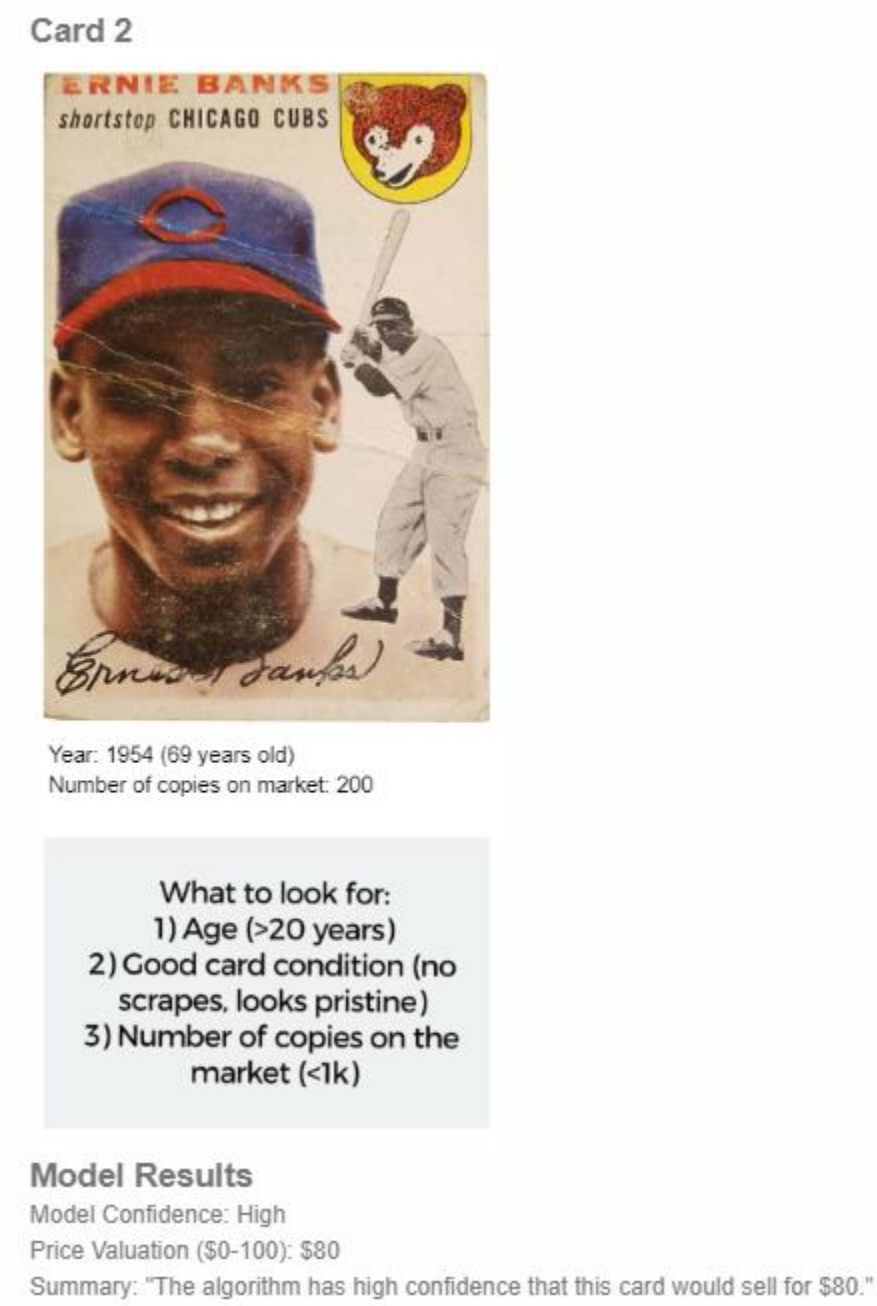


Shown above are the influences of features on the price valuation from the average ($50)

Model Results

Model Confidence: High
Price Valuation ($0-100): $80
Summary: "The algorithm has high confidence that this card would sell for $80. This card's old age increases its price $40 above average, but a few scrapes as highlighted by the algorithm's image explanation decrease its price by $30. However, since there are only 200 copies on the market, this card's valuation increases by $20."

Card II, with no model explanations:



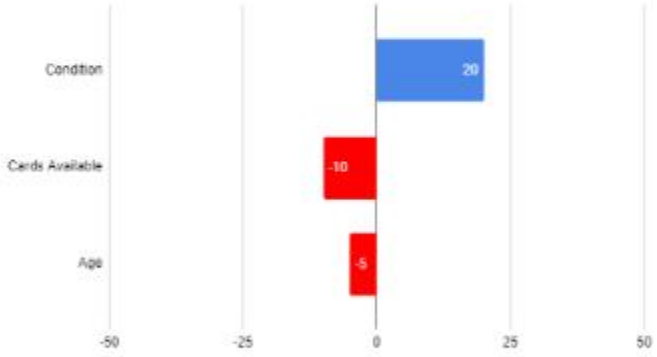
Card III, with model explanations:

Card 3

Year: 2013 (10 years old)
Number of copies on market: 3000

Shown above is the part of the image that the algorithm is recognizing to make its judgment on condition

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)



Shown above are the influences of features on the price valuation from the average ($50)

Model Results

Model Confidence: High
Price Valuation ($0-100): $55
Summary: "The algorithm has high confidence that this card would sell for $55. This card's condition increases its price $20 above average, but the algorithm's image explanation only highlights the logo on the card to make this judgment. Since the card is fairly new and there are many copies available, this decreases the card's valuation by $15."

Card III, with no model explanations:

Card 3

Year: 2013 (10 years old)
Number of copies on market: 3000

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)

Model Results

Model Confidence: High
Price Valuation ($0-100): $55
Summary: "The algorithm has high confidence that this card would sell for $55."

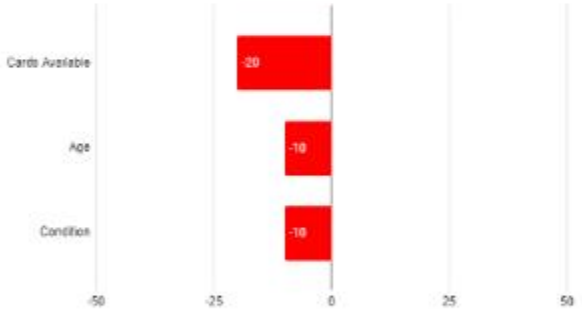
Card IV, with model explanations:

Card 4

Year: 2020 (3 years old)
Number of copies on market: 5000

Shown above is the part of the image that the algorithm is recognizing to make its judgment on condition

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)



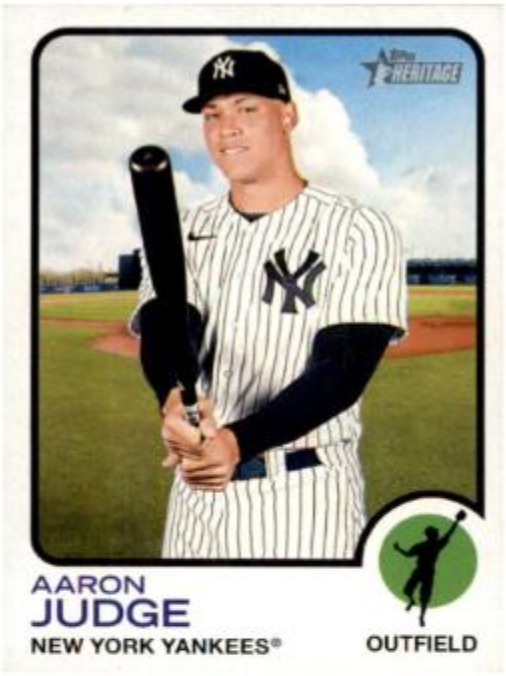
Shown above are the influences of features on the price valuation from the average ($50)

Model Results

Model Confidence: High
Price Valuation ($0-100): $10
Summary: "The algorithm has high confidence that this card would sell for $10. Since the card is new, and there are many in circulation, the valuation for this card decreases by $30 below average between the two categories. Additionally, the algorithm's image explanation highlights a few spots on the card where it claims that there are scratches; however, this does not appear to be the case. Nonetheless, the algorithm penalizes the card's valuation by $10 for its condition."

Card IV, with no model explanations:

Card 4



Year: 2020 (3 years old)
Number of copies on market: 5000

What to look for:
1) Age (>20 years)
2) Good card condition (no scrapes, looks pristine)
3) Number of copies on the market (<1k)

Model Results

Model Confidence: High
Price Valuation ($0-100): $10
Summary: "The algorithm has high confidence that this card would sell for $10."

Exhibit 15 - Scenario I, Candidate I

Exhibit 15, A

Q: Would you recommend hiring this candidate?

Explainable	Value	P-value
Yes	1 (0.78%)	
No	127 (99.22%)	< 0.001****
Total	128	

Unexplainable	Value	P-value
Yes	4 (3.15%)	
No	123 (96.85%)	< 0.001****
Total	127	

Response	Explainable	Unexplainable	P-value
Yes	127 (99.22%)	123 (96.85%)	0.181
Total	128	127	

Exhibit 15, B

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	50 (39.06%)	75 (59.06%)	0.001***
I balanced my decision with information from the algorithm	74 (57.81%)	48 (37.80%)	0.001***
I defaulted to the algorithm's recommendation	4 (3.13%)	4 (3.15%)	0.500

Total	128	127	

Exhibit 15, C

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made the correct hiring recommendation	89.88	89.80	0.961
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example	80.59	81.27	0.774
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	64.17	64.82	0.774

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable (Mean)	P-value (greater than entry below it)	Unexplainable (Mean)	P-value (greater than entry below it)
Please rate your confidence that you made the correct hiring recommendation	89.88	< 0.001****	89.80	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example	80.59	< 0.001****	81.27	< 0.001****

Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	64.17		64.82	

Exhibit 15, D

Q: Of the model explanations provided, which one helped you more in making your recommendation?

Explainable	Value	P-value
Image highlighting	45 (35.16%)	
Feature influence bar chart	83 (64.84%)	0.005***
Total	128	

Exhibit 16 - Scenario I, Candidate II

Exhibit 16, A

Q: Would you recommend hiring this candidate?

Explainable	Value	P-value
Yes	57 (44.53%)	
No	71 (55.47%)	0.108
Total	128	

Unexplainable	Value	P-value
Yes	44 (34.65%)	
No	83 (65.35%)	< 0.001****
Total	127	

Response	Explainable	Unexplainable	P-value

No	71 (55.47%)	83 (65.35%)	0.069
Total	128	127	

Exhibit 16, B

Q: Of the model explanations provided, which one helped you more in making your recommendation?

Explainable	Value	P-value
Image highlighting	32 (25%)	
Feature influence bar chart	96 (75%)	< 0.001****
Total	128	

Exhibit 16, C

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	50 (39.06%)	56 (44.09%)	0.246
I balanced my decision with information from the algorithm	75 (58.59%)	68 (53.54%)	0.246
I defaulted to the algorithm's recommendation	3 (2.34%)	3 (2.36%)	0.500
Total	128	127	

Exhibit 16, D

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable, Candidate I (Mean)	Explainable, Candidate II (Mean)	P-value
Please rate your confidence that you made the correct hiring	89.88	73.66	< 0.001****

recommendation			
Please rate your confidence in the algorithm’s ability to make the correct hiring recommendation for this example	80.59	56.97	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	64.17	59.80	0.023*

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Unexplainable, Candidate I (Mean)	Unexplainable, Candidate II (Mean)	P-value (drop)
Please rate your confidence that you made the correct hiring recommendation	89.80	75.59	< 0.001****
Please rate your confidence in the algorithm’s ability to make the correct hiring recommendation for this example	81.27	52.96	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	64.82	58.57	0.003**

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made the correct hiring	73.66	75.59	0.310

recommendation			
Please rate your confidence in the algorithm’s ability to make the correct hiring recommendation for this example	56.97	52.96	0.141
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	59.80	58.57	0.573

Exhibit 17 - Scenario I, Candidate III

Exhibit 17, A

Q: Would you recommend hiring this candidate?

Explainable	Value	P-value
Yes	106 (82.81%)	< 0.001****
No	22 (17.19%)	
Total	128	

Unexplainable	Value	P-value
Yes	88 (69.29%)	< 0.001****
No	29 (30.71%)	
Total	127	

Response	Explainable	Unexplainable	P-value
Yes	106 (82.81%)	88 (69.29%)	0.009***
Total	128	127	

Exhibit 17, B

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	103 (80.47%)	89 (70.08%)	0.038*
I balanced my decision with information from the algorithm	25 (19.53%)	36 (28.35%)	0.066
I defaulted to the algorithm's recommendation	0 (0%)	2 (1.57%)	0.237
Total	128	127	

Exhibit 17, C

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident)

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made the correct hiring recommendation	82.76	76.79	< 0.001****
Please rate your confidence in the algorithm’s ability to make the correct hiring recommendation for this example	25.89	45.82	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	45.27	54.31	< 0.001****

Exhibit 17, D

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable, Candidate II (Mean)	Explainable, Candidate III (Mean)	P-value
Please rate your confidence that you made the correct hiring recommendation	73.66	82.76	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example	56.97	25.89	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	59.80	45.27	< 0.001****

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Unexplainable, Candidate II (Mean)	Unexplainable, Candidate III (Mean)	P-value
Please rate your confidence that you made the correct hiring recommendation	75.59	76.79	0.261
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example	52.96	45.82	0.004***
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	58.57	54.31	0.029*

Exhibit 18 - Scenario I, Candidate IV

Exhibit 18, A

Q: Would you recommend hiring this candidate?

Explainable	Value	P-value
Yes	46 (35.94%)	
No	82 (64.06%)	0.001****
Total	128	

Unexplainable	Value	P-value
Yes	28 (22.05%)	
No	99 (77.95%)	< 0.001****
Total	127	

Response	Explainable	Unexplainable	P-value
No	82 (64.06%)	99 (77.95%)	0.011*
Total	128	127	

Exhibit 18, B

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	86 (67.19%)	68 (53.54%)	0.018*
I balanced my decision with information from the algorithm	40 (31.25%)	56 (44.09%)	0.023*
I defaulted to the algorithm's recommendation	2 (1.56%)	3 (2.36%)	0.497

Total	128	127	

Exhibit 18, C

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident)

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made the correct hiring recommendation	72.75	76.24	0.044*
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example	35.77	60.83	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	41.48	54.62	< 0.001****

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable, Candidate III (Mean)	Explainable, Candidate IV (Mean)	P-value
Please rate your confidence that you made the correct hiring recommendation	82.76	72.75	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example	25.89	35.77	< 0.001****
Please rate your confidence in the algorithm's ability to	45.27	41.48	0.084

make the correct hiring recommendation in general			

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Unexplainable, Candidate III (Mean)	Unexplainable, Candidate IV (Mean)	P-value
Please rate your confidence that you made the correct hiring recommendation	76.79	76.24	0.387
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation for this example	45.82	60.83	< 0.001****
Please rate your confidence in the algorithm's ability to make the correct hiring recommendation in general	54.31	54.62	0.446

Exhibit 19 - Scenario I, End of Survey Results

Exhibit 19, A

Q: Please answer the following:

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your knowledge of artificial intelligence	41.54	44.02	0.325

Exhibit 19, B

Q: Please answer the following:

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate the degree in which the algorithm influenced your decision making	41.74	35.46	0.014*

Exhibit 19, C

Q: Please answer the following:

Response (Explainable)	Bottom-two box (1, 2)	Top-two box (4, 5)	P-value
I had enough information to make hiring recommendations with confidence.	41 (32.38%)	70 (55.12%)	< 0.001****
Model explanations helped me better understand the strengths and weaknesses of the algorithm.	19 (14.96%)	79 (62.20%)	< 0.001****
Model explanations helped me better understand the strengths and weaknesses of the algorithm.	17 (13.39%)	85 (66.93%)	< 0.001****
Model explanations offered too much information and confused me in the decision-making process.	100 (78.74%)	10 (7.87%)	< 0.001****
The model explanations provided were sufficient for resolving sources of confusion.	58 (45.67%)	30 (23.62%)	< 0.001****

Response (Unexplainable)	Bottom-two box (1, 2)	Top-two box (4, 5)	P-value

I had enough information to make hiring recommendations with confidence.	43 (33.86%)	62 (48.81%)	0.011*

Exhibit 20 - Scenario I, Open-ended responses

Exhibit 20, A

Cause of errors, explainable participant group

- "For image highlighting, I think that the model was unable to capture all of the candidates' features as there were some who were smiling but the model did not detect that."
- "It did not accurately assess the candidates that were smiling which surprised me because it should have been facial recognition software."
- "Most of the AI photos did not correctly capture the person's smile."

Cause of errors, unexplainable participant group

- "With one of the clear-cut requirements being 5 years experience, it surprised me that the algorithm recommended to hire someone with less than that, especially given the numerical and objective nature of that requirement."
- "It seemed like candidates had what they were looking for, but the algorithm recommended they not be hired because of their years of experience."
- "The years of experience was a large deciding factor"
- "The algorithm was biased heavily towards the amount of work experience that the candidates had."

Exhibit 20, B

Improvement suggestions, explainable participant group

- “The algorithm should get better at picking parts of the images to analyze for specific things like smiling”
- “It should be able to detect smiling on all races if that is going to be taken into account for hiring.”
- “More testing of the algorithm, especially computer vision (smiling metric was off)”
- “bias prevention”

Improvement suggestions, unexplainable participant group

- “I feel as if the third candidate was a better option than the 4th, but the algorithm gave him the smallest percentage.”
- “The 3rd candidate should have had a better score in my opinion.”
- “they gave a 30% to the person who i recommended and i think it was only based off on the fact he had 3 instead of 5 years of experience. that was a bad call on their part because he checked off everything else “

Exhibit 20, C

Lack of trust in the algorithm, explainable group

- “There were times where the image highlighting feature did not work as intended. This could impact workers who only follow the Ai's recommendation”
- “the AI wasn't accurately processing where the candidate's face and smile was, therefore allowing me to not trust it at all “
- “There were two instances where the AI could not identify the candidate's face and therefore gave them a poor smile rating, which led the AI to decide that otherwise worthy candidates did not deserve the job. I certainly wouldn't trust this AI to make hiring decisions for the HR department of a company I manage, and I wouldn't take the recommendations into consideration when reviewing applicants.”

- “I was surprised by the model explanation that it failed to capture some of the candidates smiling. It made the AI feature less credible as a beneficiary for the hiring process.”

Lack of trust in the algorithm, unexplainable group

- “I could not predict what the algorithm was going to recommend and it made me lose trust in the algorithm's decisions.”
- “The algorithm seemed to often disagree with my thinking.”
- “Sometimes I thought a candidate was perfect but the algorithm said they would not make a good hire”

Exhibit 20, D

Desire for more information, explainable group

- “A lot more clarity in regards to years of experience”
- “The AI should be able to give more human-like reasoning for how it arrived at its conclusion, including real internal compromises, instead of using its flawed analyses to make judgments based on "balancing out" strengths and weaknesses.”
- “I think more in-depth explanations could be helpful; the more information provided the better.”
- “The model explanations could simply provide more detailed information, or account for any AI errors.”
- “The model was very bad at identifying smiles and also I don't understand how it was matching up synonyms.”

Desire for more information, unexplainable group

- “I would like to know what weight was attached to the years of experience vs. the qualities of the candidates. This would help me understand why the algorithm arrived at the decisions that it did.”

- “I would have liked to know the scoring process of each requirement and the weightings of each, so I could better understand the algorithm's process of getting to its final score.”
- “How is "hire"/"do not hire" calculated? Why does the algorithm approve candidates with less than 5 years of experience?”
- “I would have liked to know how the algorithm made its decisions -- namely, what was the cutoff for whether or not the algorithm recommended a candidate? How much was each criterion weighted? These would have helped me better understand the algorithm's recommendations and probably would have helped me develop more trust in it. “
- “I would have liked to know how the algorithm makes it decisions; what inputs it takes and how it processes them”

Exhibit 21 - Scenario II, Card I

Exhibit 21, A

Q: Would you recommend listing this card for $70?

Explainable	Value	P-value
Yes	75 (58.14%)	0.032*
No	54 (41.86%)	
Total	129	

Unexplainable	Value	P-value
Yes	62 (48.82%)	
No	65 (51.18%)	0.395
Total	127	

Response	Explainable	Unexplainable	P-value
Yes	75 (58.14%)	62 (48.82%)	0.160

Total	129	127	

Q: If you said “no” above, please set your card price below:

Response	Explainable (Mean)	P-value (against price)	Unexplainable (Mean)	P-value (against price)	P-value (difference between groups)
Price	$56.67	< 0.001****	$54.62	< 0.001****	0.529
Total	55		71		

Exhibit 21, B

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	15 (11.63%)	20 (15.75%)	0.218
I balanced my decision with information from the algorithm	102 (79.07%)	102 (80.31%)	0.463
I defaulted to the algorithm's recommendation	12 (9.30%)	5 (3.94%)	0.070
Total	129	127	

Exhibit 21, C

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made an appropriate valuation	68.29	73.87	0.004***
Please rate your confidence in the	68.69	67.77	0.349

algorithm’s ability to make an appropriate valuation for this example			
Please rate your confidence in the algorithm's ability to make appropriate valuations in general	67.60	69.56	0.170

Exhibit 21, D

Q: Of the model explanations provided, which one helped you more in making your recommendation?

Explainable	Value	P-value
Image highlighting	35 (27.13%)	
Feature influence bar chart	94 (72.87%)	< 0.001****
Total	129	

Exhibit 22 - Scenario II, Card II

Exhibit 22, A

Q: Would you recommend listing this card for $80?

Explainable	Value	P-value
Yes	54 (41.86%)	
No	75 (58.14%)	0.032*
Total	129	

Unexplainable	Value	P-value
Yes	45 (35.43%)	
No	82 (64.57%)	< 0.001****
Total	127	

Response	Explainable	Unexplainable	P-value
No	75 (58.14%)	82 (64.57%)	0.177
Total	129	127	

Q: If you said “no” above, please set your card price below:

Response	Explainable (Mean)	P-value (against price)	Unexplainable (Mean)	P-value (against price)	P-value (difference between groups)
Price	$80.84	0.625	$79.10	0.370	0.625
Total	81		90		

Exhibit 22, B

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	26 (20.16%)	36 (28.35%)	0.083
I balanced my decision with information from the algorithm	93 (72.09%)	83 (65.35%)	0.152
I defaulted to the algorithm's recommendation	10 (7.75%)	8 (6.30%)	0.417
Total	129	127	

Exhibit 22, C

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made an appropriate	75.14	79.28	0.019*

valuation			
Please rate your confidence in the algorithm’s ability to make an appropriate valuation for this example	66.02	70.30	0.040*
Please rate your confidence in the algorithm's ability to make appropriate valuations in general	68.97	69.42	0.417

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable, Card I (Mean)	Explainable, Candidate II (Mean)	P-value
Please rate your confidence that you made an appropriate valuation	68.29	75.14	< 0.001****
Please rate your confidence in the algorithm’s ability to make an appropriate valuation for this example	68.69	66.02	0.124
Please rate your confidence in the algorithm's ability to make appropriate valuations in general	67.60	68.97	0.255

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Unexplainable, Card I (Mean)	Unexplainable, Card II (Mean)	P-value
Please rate your confidence that you made an appropriate valuation	73.87	79.28	0.003***

Please rate your confidence in the algorithm's ability to make an appropriate valuation for this example	67.77	70.30	0.155
Please rate your confidence in the algorithm's ability to make appropriate valuations in general	69.56	69.42	0.474

Exhibit 23 - Scenario II, Card III

Exhibit 23, A

Q: Would you recommend listing this card for $55?

Explainable	Value	P-value
Yes	21 (16.28%)	
No	108 (83.72%)	< 0.001****
Total	129	

Unexplainable	Value	P-value
Yes	23 (18.11%)	
No	104 (81.89%)	< 0.001****
Total	127	

Response	Explainable	Unexplainable	P-value
No	108 (83.72%)	104 (81.89%)	0.412
Total	129	127	

Q: If you said "no" above, please set your card price below:

Response	Explainable (Mean)	P-value (against price)	Unexplainable (Mean)	P-value (against price)	P-value (difference between groups)
Price	$36.33	< 0.001****	$34.55	< 0.001****	0.306
Total	109		110		

Exhibit 23, B

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	59 (45.74%)	60 (47.24%)	0.454
I balanced my decision with information from the algorithm	62 (48.06%)	60 (47.24%)	0.498
I defaulted to the algorithm's recommendation	8 (6.20%)	7 (5.51%)	0.500
Total	129	127	

Exhibit 23, C

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made an appropriate valuation	71.81	74.46	0.096
Please rate your confidence in the algorithm’s ability to make an appropriate valuation for this example	55.40	56.91	0.301
Please rate your	62.25	65.55	0.378

confidence in the algorithm's ability to make appropriate valuations in general			

Exhibit 24 - Scenario II, Card IV

Exhibit 24, A

Q: Would you recommend listing this card for $10?

Explainable	Value	P-value
Yes	57 (44.19%)	
No	72 (55.81%)	0.093
Total	129	

Unexplainable	Value	P-value
Yes	87 (68.50%)	< 0.001****
No	40 (31.50%)	
Total	127	

Response	Explainable	Unexplainable	P-value
No	72 (55.81%)	40 (31.50%)	0.025*
Total	129	127	

Q: If you said “no” above, please set your card price below:

Response	Explainable (Mean)	P-value (against price)	Unexplainable (Mean)	P-value (against price)	P-value (difference between groups)
Price	$20	< 0.001****	$17.42	0.004**	0.297
Total	80		55		

Exhibit 24, B

Q: How did you arrive at your recommendation?

Response	Explainable	Unexplainable	P-value
I made this decision on my own	36 (27.91%)	21 (16.54%)	0.021*
I balanced my decision with information from the algorithm	76 (58.91%)	76 (59.84%)	0.491
I defaulted to the algorithm's recommendation	17 (13.18%)	30 (23.62%)	0.023*
Total	129	127	

Exhibit 24, C

Q: Please rate your confidence for the following (0-100, 0 = not confident, 100 = very confident):

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your confidence that you made an appropriate valuation	74.32	78.29	0.037*
Please rate your confidence in the algorithm’s ability to make an appropriate valuation for this example	61.15	73.31	< 0.001****
Please rate your confidence in the algorithm's ability to make appropriate valuations in general	63.07	69.13	0.007***

Exhibit 25 - Scenario II, End of Survey Results

Exhibit 25, A

Q: Please answer the following:

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate your knowledge of artificial intelligence	39.66	46.64	0.006***
Please rate your knowledge of baseball	35.30	36.33	0.392

Exhibit 25, B

Q: Please answer the following:

Response	Explainable (Mean)	Unexplainable (Mean)	P-value
Please rate the degree in which the algorithm influenced your decision making	60.91	55.78	0.018*

Exhibit 25, C

Q: Please answer the following:

Response (Explainable)	Bottom-two box (1, 2)	Top-two box (4, 5)	P-value
I had enough information to make card valuations with confidence.	24 (18.75%)	69 (53.91%)	< 0.001****
Model explanations helped me better understand the strengths and weaknesses of the algorithm.	5 (3.90%)	109 (85.16%)	< 0.001****
Model explanations helped me better understand the strengths and weaknesses of the algorithm.	12 (9.38%)	87 (67.97%)	< 0.001****
Model explanations offered too much	113 (88.28%)	4 (3.13%)	< 0.001****

information and confused me in the decision-making process.			
The model explanations provided were sufficient for resolving sources of confusion.	25 (19.53%)	57 (44.53%)	< 0.001****

Q: Please answer the following:

Response (Unexplainable)	Bottom-two box (1, 2)	Top-two box (4, 5)	P-value
I had enough information to make card valuations with confidence.	30 (23.62%)	69 (54.33%)	< 0.001****

Exhibit 26 - Scenario II, Open-ended responses

Exhibit 26, A

Cause of errors, explainable group

- "The image highlighting wasn't always the most accurate (ex: highlighting just the logo and scuffs/flaws that weren't there)."
- "The image highlighting seemed a bit sketchy for a few examples, like for example on the last card it spotted scratches that didn't seem to exist."
- "The image highlighting prompted detections that weren't necessarily accurate"
- "Sometimes it didn't seem that highlighted parts of the card were damaged"

Cause of errors, unexplainable group

- "The algorithm made fairly modest valuations."
- "It seemed to make conservative results that stayed closer to the average price than I personally would've priced the cards at."

- “The old baseball card was in horrible condition and still valued at $80.”

Exhibit 26, B

Debugging procedures, explainable group

- “The card that image highlighted the logo surprised me because there did not seem to be anything wrong with that card”
- “I noticed that the image highlighting was sometimes inaccurate when the cards were in pretty good condition. This could make pricing lower than market value.”
- “It seemed to focus on specific aspects like logos”

Debugging procedures, unexplainable group

- Card 3 was a basic card, yet the algorithm was trying to sell it for above average. I was shocked to see Card 4 be priced so low considering the similarities between the features of the cards.”
- “The algorithm's prices were extremely close to the prices I already had in mind, meaning it was evaluating the cards very similarly to me.”
- “I thought that the algorithm was more accurate for the first 2 cards, but less accurate for the last 2.“

Exhibit 26, C

Confusion around model explanations, explainable group

- “The model was often making decisions on partial images which clearly impacted the value of the card. So the valuations were being skewed in a way I wasn't expecting.”
- “Didn't completely understand why algorithm couldn't read the entire image, also how to quantify something like condition”
- “the AI system only picked up on parts of the card so it seemed to be looking at an inaccurate representation when analyzing “

- "The bar chart was a bit hard to read at first so it took me longer to understand/comprehend the data being presented."

Exhibit 26, D

Desire for more information, explainable group

- I was surprised that the model explanations did not include specific reasonings for deductions"
- "Image highlighting was not consistent over different cards. I believe that it should be consistent. Also, the evaluation of a card's condition is rather subjective and it is not clear which exact parameters are used by the model to give grades."
- "I would want reasoning as to why specific amounts of deductions were issued"
- "I would like to see an example of an average card to compare"
- "I think the explanations are fine, but you really need the model information to give it some needed context. "
- "Comparable valuations, previously sold cards, market conditions."
- "Perhaps more explanation for why certain parts of the image are highlighted versus others."

Desire for more information, unexplainable group

- "It did not give tons of information"
- "The algorithm seemed pretty accurate, although it was hard to tell."
- "Was it equally geared towards all three factors? Or was one of the three factors of more importance?"
- "How much did each piece of criteria (age, number of cards available, etc.) factor into the price? What was each categories worth?"
- "I wanted to know how the algorithm evaluated the condition of the card."
- "I would love to know how each factor was weighted, by how much, and whether an increase/decrease in any factor was weighed exponentially."

- “I would have liked to know what numerical values were assigned to each criteria to get a better understanding of how the algorithm arrived at each price.”
- “I would like to know if the algorithm factored in prices of similar cards found online into the equation.”
- “I would assume that there is an equation with just three variables that determines the price. Would like to know the price function to understand the different weights associated with each variable.”