Journal of the Academy of Marketing Science (2020) 48:137–141
https://doi.org/10.1007/s11747-019-00710-5

COMMENTARY

Explainable AI: from black box to glass box

Arun Rai[1]

Published online: 17 December 2019

Introduction

This special issue profiles how technological innovations are exerting a transformative force on the practice and academic discipline of marketing. These technologies create tremendous upside potential along with novel risks that need to be understood and effectively managed to realize the benefits and mitigate the downsides. Each of the technological advances discussed in the special issue—health IT (Agarwal et al.), robotics (Davenport et al.), chatbots (Thomaz et al.), mobile (Tong et al.), social media (Appel et al.), and in-store retail technology (Grewal et al.), is fueled in significant ways by artificial intelligence (AI).

The torrent in the development and deployment of AI systems is expanding the scale and scope with which these systems are affecting our work and everyday lives. These systems are penetrating a broad range of industries, such as education, construction, healthcare, news and entertainment, travel and hospitality, logistics, manufacturing, law enforcement, and finance. Their role is becoming much more profound in our lives by influencing what we buy, who we hire, who our friends are, what newsfeed we receive, and even how our children and elderly are cared for. They are being employed in a range of marketing applications, such as personalizing product and content recommendations and optimizing cost-per-click and cost-per-acquisition in ad targeting by mining troves of online consumer behavior data. Frontier applications are predicting individuals' future needs and recommending actions to them. For example, Amazon's recently launched add-on feature to its personal assistant Alexa, "Alexa Hunches", learns individual's rhythms in interacting with smart-home devices such as a lock or door, observes deviance from rhythms, and reminds users when to lock a door or turn off a light.

However, we typically have little understanding on why AI systems make decisions or exhibit certain behaviors. Many machine learning (ML) algorithms used to develop these systems are inscrutable, particularly deep learning neural network approaches which have emerged to be a very popular class of ML algorithms. This inscrutability can hamper users' trust in the system, especially in contexts where the consequences are significant, and lead to the rejection of the systems. It has also obfuscated the discovery of algorithmic biases arising from flawed generated processes that are prejudicial to certain groups. Such biases have led to large-scale discrimination based on race and gender in a number of domains ranging from hiring to promotions and advertising to criminal justice to healthcare. Such biases against vulnerable populations in healthcare are discussed in Agarwal et al. in this issue, while Davenport et al. (also in this issue) provide a broader discussion of algorithmic biases.

The following definition of AI makes salient that human users need to trust AI systems in attaining objectives (Russell n.d, p. 11): "Machines are *beneficial* to the extent that *their* actions can be expected to achieve *our* objectives." What should be the basis for this trust? In addition to providing users with information on the system's prediction accuracy and other facets of performance, providing users with an effective explanation for the AI system's behavior can enhance their trust in the system. In situations where the system makes a recommendation to the user on a product to purchase or a connection to add to their professional or social network, an explanation for the recommendation is likely to make the information more useful to the user and have a stronger influence on the user's actions. Such explanations can also be leveraged by developers to improve the model through feature engineering, modification of the model's architecture, and tuning of hyperparameters, and by trainers to revise the set of learning and testing data resources.

Explainable AI (XAI) is the class of systems that provide visibility into how an AI system makes decisions and

✉ Arun Rai
arunrai@gsu.edu

[1] J. Mack Robinson College of Business, Georgia State University, 35 Broad Street NW, Atlanta, GA, USA

predictions and executes its actions. XAI explains the rationale for the decision-making process, surfaces the strengths and weaknesses of the process, and provides a sense of how the system will behave in the future.[1]

Given the extensiveness with which AI systems are being developed and deployed to upend marketing, as illustrated in the articles in the special issue—from personalizing the user experience, to recommending products and content for customers, to lead-scoring for B2B marketing teams, to automating two-way conversations with customers to nurture relationships—it becomes critical for marketing researchers to understand how to achieve explainability for different types of AI models, assess the tradeoff between prediction accuracy and explanation associated with different choices, and develop and deploy trustworthy AI systems that meet business and fairness objectives.

I briefly differentiate between inherently interpretable AI models and black-box deep learning models, overview XAI approaches to turn black-box models into glass-box models, and discuss the research implications related to leveraging XAI in marketing AI applications.

Inherently interpretable models vs. black-box deep-learning models

The process to generate explanations underlying the behavior of AI systems will depend on the type of ML algorithms: algorithms that generate inherently interpretable models versus deep learning algorithms that are complicated in structure and learning mechanisms and generate models that are inherently uninterpretable to human users (Hall and Gill 2019; Du et al. 2018).

Machine learning algorithms such as decision trees, Bayesian classifiers, additive models, and spare linear models generate interpretable models in that the model components (e.g., weight of a feature in a linear model, a path in a decision tree, or a specific rule) can be directly inspected to understand the model's predictions. These algorithms use a reasonably restricted number of internal components (i.e. paths, rules, or features) but provide traceability and transparency in their decision making. As long as the model is accurate for the prediction task, these approaches provide the visibility to understand decisions made by the AI system.

In contrast, deep learning algorithms are a class of ML algorithms which sacrifice transparency and interpretability for prediction accuracy. These algorithms are now being employed to develop applications such as prediction of consumer behaviors based on high-dimensional inputs, speech recognition, image recognition, and natural language processing. As an example, convolutional neural networks, which underlie facial recognition applications, extract high-level complex abstractions of a face through a hierarchical learning process which transforms pixel-level inputs of an image to relevant facial features to connected features that abstract to the face. The model learns the features that are important by itself instead of requiring the developer to select the relevant feature. As the model involves pixel-level inputs and complex connections across layers of the network which yield highly nonlinear associations between inputs and outputs, the model is inherently uninterpretable to human users.

Addressing the trade-off between prediction and explanation associated with deep learning models, there have been significant recent advances in *post-hoc interpretability techniques*—these techniques approximate deep-learning black-box models with simpler interpretable models that can be inspected to explain the black-box models. These techniques are referred to XAI as they turn black-box models into glass-box models and are receiving tremendous attention as they offer a way to pursue both prediction accuracy and interpretability objectives with AI applications.

Classes of XAI: How to convert black-box models to glass-box models

While there are different types of post-hoc explanation techniques for black-box models, they all require the representation of input variables to be interpretable to humans. For example, while a deep learning text classifier of customer sentiment uses complex features such as word embeddings which are uninterpretable to users or developers, an interpretable representation may be a binary vector indicating the presence or absence of words. As such, the inputs to post-hoc explanation techniques may need to differ from the inputs to deep learning models.

A number of classifications of XAI techniques for deep learning models have been proposed (e.g., Hall and Gill 2019; Du et al. 2018; Ribeiro et al. 2016). Drawing on this work, XAI techniques can be classified using two dimensions: (i) whether the technique is model-specific or model-agnostic and (ii) whether the technique is designed to provide an explanation that is global in scope to the model or one that is local in scope to a prediction (Table 1).

Model-specific techniques incorporate interpretability constraints within the inherent structure and learning mechanisms underlying deep learning models, whereas model-agnostic techniques use the inputs and predictions of the black box models to generate explanations.

The scope of the explanation, global to the model versus local to the prediction, corresponds to trust at two levels—

[1] The Defense Advanced Research Projects Agency (DARPA) program on XAI identifies these as key characteristics of XAI: Turek, Matt, Explainable AI, Program Information, Defense Advanced Research Projects Agency, https://www.darpa.mil/program/explainable-artificial-intelligence, Accessed on October 20, 2019.

Table 1 Classification of XAI Techniques

	Model-specific	Model-agnostic
Global	Enforce interpretability constraints into the structure and learning mechanisms of deep learning models	Develop interpretable global surrogate models based on input-output associations predicted by a black-box model Apply diagnostic techniques to understand the importance of specific features in a black-box model's predictions
Local	Use attention mechanisms to show how the model selectively focuses on features in high-dimensional input for an instance	Develop interpretable surrogate models with local fidelity in the vicinity of an instance

trust in the model versus *trust in the prediction* (Ribeiro et al. 2016). Explanations about how the black-box model makes its predictions can influence the trust of users and developers in the model and consequently the confidence they have in deploying the model. Explanation about a prediction can influence a user's *trust in the prediction* and consequently whether or not the user takes an action based on the prediction.

I now briefly review each of the classes of XAI techniques along with illustrative examples.

Model-specific global explanation

These XAI techniques increase the comprehensibility of the models by incorporating interpretability constraints into the structure of the model. The structural constraints can include: *sparsity* (where fewer features are used as inputs) and *monotonicity* (where the relationship between features and predictions is constrained to be monotonic). Semantic meaningfulness constraints can also be incorporated to limit the higher-level abstractions that are extracted from the data; for example, a convolutional neural network for facial recognition can be constrained to learn disentangled representations such as forehead, eyes, nose, cheeks, and lips that are comprehensible to the user. By doing so, the system can detect and pool information across parts of the face to differentiate between types of emotions including Happy, Sad, Angry, Surprised, Disgusted, Calm, Confused and Fear.

Model-specific local explanation

This class of XAI techniques provides an explanation for a specific instance in a deep learning model. A simple example would be surfacing the input features pertaining to a customer's purchasing history or social networks that led a deep learning model to target a specific ad to a customer in a given spatiotemporal context such as when the customer is proximate to a store location on a given day and time. Such an approach can also be used to understand the role of hyper-context features in a deep learning model for high-precision mobile targeting strategies, such as those discussed in Tong et al. in this issue.

Attention mechanisms, a novel technique in this category, highlight for users the importance of different parts of a high-dimensional input stream that are the basis for a model's automated description of an instance. Consider a deep learning model which uses a convolutional neural network to encode an image to a vector and a recurrent neural network that uses this vector to generate a caption for the image. An attention mechanism scheme can be employed with the recurrent neural network to show to the user what image segments the model focuses on to generate each substantive word of the caption for the image (Xu et al. 2015). For example, for an image captioned by a deep learning model as *Customer* is *confused* by the *choice of products*, the attention scheme would show the relevant segments of the image corresponding to the italicized words.

Model-agnostic global explanation

With this class of XAI techniques, the explanation process involves approximating an interpretable model for the black-box model. For example, a deep learning model on how location-aware mobile advertising affects customer response can be approximated with an interpretable decision tree. The IF-THEN logic of the decision tree can provide an explanation of the relative importance of factors in influencing customer response to the mobile advertisement. Domain experts can inspect these factors and are likely to trust the model to the extent that the factors are deemed to be reasonable and not to be conflating noise as signal.

Diagnostic techniques can be employed to generate insights on the importance of specific features in the model's predictions. Partial dependence plots can be used to assess the marginal effects of selected features on the prediction outcomes, while individual conditional expectation can be used to gain a granular view on how a feature impacts individual instances and discover heterogeneity in impacts across instances. For example, a partial dependence plot can illuminate the importance of customer's emotions, as detected by a facial recognition system, in responding favorably to in-store promotions; in contrast, individual conditional expectation can surface the heterogeneity in this impact across microsegments of customers.

Model-agnostic local explanation

The objective with this category of XAI techniques is to generate model-agnostic explanations for a specific instance or for the vicinity of a specific instance.

A recently developed technique, Local Interpretable Model-Agnostic Explanation (LIME), develops an explanation of a model's behavior in the *neighborhood of an instance* (Ribeiro et al. 2016). Consider a deep learning model that classifies a product to be at high risk for significant decline in sales for which a post-hoc explanation is sought by the product manager. The interpretable components (e.g., trend in sales, recent and expected competitive moves, various sentiments that are expressed by customers in customer-service chat sessions, social-media engagement with the product brand) are perturbed to evaluate how the predictions made by the classifier change. A linear model is learned on this perturbed dataset with greater weights accorded to the perturbated instances in the vicinity of the product. The components of the linear model with the greatest importance that suggest the product sales are at high risk or the converse are identified as an explanation for the classifier and can be rendered to the product manager in an understandable manner. A similar process can be employed to understand why a facial recognition system classifies a customer as Confused or Angry, each of which would require a very different customer service response. With social media actively deployed for marketing (see Appel et al. in this issue), this process can also be employed to understand the role of interpretable facets of social media content from text, images, likes, and emoticons in posts and comments explains customer engagement with an ad or a brand.

Implications of XAI for marketing research

Developing and deploying AI systems for marketing requires that these systems provide users with explanations that are faithful to the model and intelligible to them, while maintaining high levels of prediction accuracy. How to realize the potential of XAI approaches to address the tension between prediction and explainability underlying black-box deep-learning models opens up exciting research avenues for marketing scholars as discussed below:

Understanding the prediction accuracy-explainability state space

What are the desired, acceptable, and unacceptable thresholds of the prediction-accuracy and explainability state space for different marketing AI applications? How do deep learning techniques, in combination with different classes of XAI approaches, perform with respect to the two competing objectives for different applications?

AI systems for highly consequential decision domains, such as pricing, promotion value and timing, product recall, and customer service, are likely to require a high level of explainability for marketing professionals and customers. In contrast, users of a personal assistant such as Alexa or Cortana that ask the assistant to tell a joke or play a song are unlikely to require much explainability. Determining the prediction accuracy vis-à-vis explainability requirements for different classes of marketing applications will enable marketing professionals to understand the objectives to be met by AI systems and to work with AI application developers to explore how best to achieve these objectives by combining XAI approaches with deep learning approaches.

Making AI trustworthy through instance-level and model-level explanations

Marketing professionals and customers who use an AI system can be skeptical of the system if they are unclear about the motives and reasonableness of the system. Their trust in the AI system can operate at two different levels—*the prediction or action* and *the model*. Research on how different XAI approaches can influence users' trust at each of these two levels in different application domains can contribute to our understanding on how explanation capabilities can influence the trust in AI applications. Studies along these lines can assess how consumers' trust in an AI-based ad targeting system can be developed by XAI which surfaces as to why specific ads are targeted to a consumer and the features underlying the ad-targeting model. Work along these lines can also assess how feedback from consumers on the reasonableness of explanations can be used to improve ad targeting and reduce the likelihood of the ads being seen as clickbait.

As AI applications are scaled to engage with customers through the end-to-end purchasing process from awareness to comparison of alternatives to post-purchase support, research can examine how XAI can be combined with the engagement process to render effective explanations at the levels of specific actions and the overall model to enhance consumer satisfaction and learning. Research can also examine how XAI can be employed to augment conversational agents so customers are able to understand, at levels that they desire, specific queries and recommendations by these agents (see Thomaz et al. in this issue for a discussion on explainability for effective interactions with conversational agents). Similar studies can also be undertaken to provide insights on how XAI can be deployed to generate specific instance-level and general model-level explanations to accompany a range of AI system's actions across domains, for example recommendations for products, content, and friends in online social networks; and personalization of newsfeed.

Achieving AI fairness

XAI techniques can be used to reveal whether attributes such as race or gender, or socio-economic and locational variables that proxy for them, are directly or indirectly used in black-

box models so the models are biased against certain groups. Research on *fairness in marketing AI* can generate insights on how XAI can be integrated with the development and deployment of AI systems to prevent and detect algorithmic bias in applications from recommendation systems to reputation scoring to targeting promotions and advertisements.

Modifying the privacy calculus

A number of marketing AI applications, such as personalized recommendations and mobile advertising, are based on the use of personal information. However, privacy is an increasingly dominant concern for consumers as discussed by Thomaz et al. in this special issue. The role of XAI in influencing the privacy calculus of individuals is therefore an important area of research. How will consumer's willingness to share their personal information change with an explanation on how the data are used and the resulting benefits for the consumer? How do explanations at the level of specific instances and the overall model affect the privacy calculus for different marketing applications?

Aligning the levels of explanation and transparency to users

Machine learning professionals may be able to inspect and interpret certain ML models and outputs generated by post-hoc interpretability techniques, but end-users are unlikely to be able to do so. Complicated explanations and high levels of transparency regarding the functioning of the AI models can impose significant attention costs, cause information overload, and frustrate users. Although a user may be interested in understanding why a social network connection is suggested, the typical user of social media applications is unlikely to be interested in understanding how network models were employed to arrive at the recommendation. A simple explanation that is accessible to the user while being faithful to the model is likely to suffice. Similarly, high levels of transparency on the underlying models are unlikely to be of interest to users, although they may be quite relevant for developers. Understanding the levels of explainability and transparency that align with the needs of different users will ensure that the appropriate levels are planned for and achieved and a one-shoe-size-fits-all approach is not pursued.

Realizing value for development vs. deployment through explainability

Designers and trainers of AI systems, internal users in an organization, and customers, can generate value from explanations on the predictions, decisions, and actions of AI systems: for example, how to engineer the features for designers; how training and test datasets may need to be adjusted for trainers; what tasks to delegate and not to delegate to AI for marketing professionals; and understanding recommendations of products and social-network connections or decisions to reject loan applications for consumers. Examining XAI utility holistically from the perspective of different stakeholders will provide a nuanced understanding about how to leverage XAI through the development and deployment lifecycle of a marketing AI application.

Concluding remarks

Advances in XAI offer ways to unmask AI black-box models and pursue two goals with AI—prediction accuracy and explanation, which have largely been treated as incompatible. Understanding how to achieve this potential opens exciting research avenues for marketing scholars on how XAI choices can redefine the prediction-accuracy and explainability tradeoff, how XAI can be leveraged to build trustworthy AI and to achieve AI fairness, how explanations on the use of personal information by algorithms can redefine the privacy calculus of consumers, and how the level of explanation and transparency can be aligned with the needs of the different stakeholders involved in the development, deployment and use of the systems.

References

Du, M., Liu, N., & Hu, X. (2018). Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033*.

Hall, P., & Gill, N. (2019). *An introduction to machine learning interpretability*. Second edition. Sebastopol, CA: O'Reilly Media, Incorporated.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM.

Russell, S. (n.d.). Human compatible: Artificial intelligence and the problem of control, Penguin Publishing Group. Kindle Edition.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R.S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057).