Research Agenda for Basic Explainable AI

Conference Paper · August 2021

CITATIONS
3

READS
710

5 authors, including:

Roman Lukyanenko
University of Virginia
181 PUBLICATIONS **1,479** CITATIONS
SEE PROFILE

Arturo Castellanos
William & Mary
44 PUBLICATIONS **280** CITATIONS
SEE PROFILE

Binny M. Samuel
University of Cincinnati
36 PUBLICATIONS **266** CITATIONS
SEE PROFILE

Monica Chiarini Tremblay
William & Mary
63 PUBLICATIONS **1,016** CITATIONS
SEE PROFILE

Research Agenda for Basic Explainable AI

Emergent Research Forum (ERF)

Roman Lukyanenko
HEC Montreal
roman.lukyanenko@hec.ca

Arturo Castellanos
Baruch College (CUNY)
Arturo.castellanos@baruch.cuny.edu

Binny M. Samuel
University of Cincinnati
samuelby@uc.edu

Monica Tremblay
College of William and Mary
monica.tremblay@mason.wm.edu

Wolfgang Maass
Saarland University
German Research Center for Artificial Intelligence (DFKI)
wolfgang.maass@dfki.de

Abstract

Artificial Intelligence is increasingly driven by powerful but often opaque machine learning algorithms. These black-box algorithms achieve high performance but are not explainable to humans in a systematic and interpretable manner, a challenge known as Explainable AI (XAI). Informed by a synthesis of two converging literature streams on information systems development and psychology, we propose a new XAI approach termed *Basic Explainable AI* and a subsequent research agenda. We propose four research directions that focus on providing explanations by proactively considering the target audience's mental models and making the explanations maximally accessible to heterogeneous nonexpert users.

Keywords

Explainable AI, machine learning, basic level categories, Basic XAI, model interpretability.

Introduction

Artificial intelligence (AI) is a preeminent technological trend of the 21st century, transforming organizations, industries, and daily lives (Anderson et al. 2018). It is estimated that AI-based analytics will contribute $15 trillion to global GDP by 2030 (Rao and Verweij 2017).

Although AI has traditionally focused on logic-based, model-driven learning, the ubiquity of data and computing power has shifted the focus towards machine learning (ML) and the creation of powerful, but often opaque, "black box" learning models (Crevier 1993; McCorduck 2004). Complex machine learning models, such as those based on *deep learning*, have further accelerated the applications of AI. The growing societal reliance on powerful (but complex) ML models creates a new societal challenge: being able to understand and explain how and why these models make their decisions – a challenge known as ***Explainable AI*** (XAI) (Castelvecchi 2016; Gunning 2016; Mueller et al. 2019).

XAI is a significant challenge that stands in the way of realizing the full benefit of AI. For example, in the absence of transparency of how AI makes decisions, decision makers and the public remain skeptical of relying on AI, especially for critical decisions and actions (Bailetti et al. 2016; Holzinger et al. 2018; Rai 2020; Sun and Medaglia 2019). The inability to understand why AI makes certain decisions may also preclude humans from detecting biases and discriminatory practices embedded in AI.

XAI is a growing research area (Dosilović et al. 2018; Mueller et al. 2019; Rai 2020) and is also being undertaken by IS scholars (Lukyanenko et al. 2020; Rai 2020). However, as recent reviews of XAI suggest, a critical gap in the area is emerging: "most of the existing literature on XAI methods are based on the developer's intuition rather than ... on the intended users" (Adadi and Berrada 2018). Furthermore, many of the approaches, as recent surveys of the area conclude, are themselves in need of further clarification and explanation (Adadi and Berrada 2018; Miller et al. 2017).

As AI becomes ubiquitous, the need for accessible explanations, especially to non-experts, grows. For example, under the European Union's "General Data Protection Regulation 2016/679," companies need to comply and provide their customers with "meaningful information about the logic involved" in their computer programs (Article 13.2(f)). The implication is that the explanations need to be made accessible to highly heterogeneous audiences, with different backgrounds, values, and beliefs.

In this paper we propose a new approach for explainable AI termed **Basic Explainable AI** or **Basic XAI**. It is informed by a synthesis of two converging literature streams on information systems (IS) development and psychology. Borrowing from the notion of *basic level categories* in psychology (Rosch et al. 1976), which was recently applied to the context of data management and interface design (Castellanos et al. 2020; Lukyanenko et al. 2019), we propose a research agenda for Basic Explainable AI, which focuses on providing explanations by making the explanations maximally accessible to heterogeneous nonexpert users.

Background: XAI and Basic Level Categories

There are several approaches to XAI. Broadly, XAI methods can be classified based on their scope as *global* (an explanation of the entire AI model) or *local* (explanations for specific decisions in the model). XAI methods can also be classified based on whether the explanations are intrinsic to a model (e.g., paths in decision trees), or provided as a post hoc supplement regardless of the model (*model-agnostic, surrogate models)* (Dosilović et al. 2018; Rai 2020). An example of an approach to XAI is the popular LIME library, which provides a local, model-agnostic explanation (Ribeiro et al. 2016). An example of a method contributed by the IS community is Superimposition, a global, model-agnostic method which enmeshes conceptual modeling diagrams with decision weights available as outputs from common machine learning models (Lukyanenko et al. 2020). The latter example is an example of using graphical notation for explanations, however these explanations could be made using natural language too (e.g., Krening et al. 2016).

As noted in the recent analysis of XAI, a common limitation of the prevailing approaches is that they do not explicitly consider the consumers of the explanation in how the approaches are developed (Abdul et al. 2018; Adadi and Berrada 2018; Miller et al. 2017; Zhu et al. 2018). This limitation is becoming more acute, as broader audiences, including nonexperts in both AI and computing, are seeking explanations from these models. This is of course not to suggest that some of the current methods are ineffective and cannot be used by nonexperts. However, the practice of XAI is in shortage of theoretically grounded and empirically validated approaches that focus on the consumers of the explanation and not on the technical feasibility or computational efficiency.

Basic XAI addresses the calls for approaching explainability from the point of view of the target audience. The motivation behind Basic XAI comes from the field of participatory design and requirements elicitation in the IS discipline. First, we observe that IS development is actively dealing with the challenge of engaging heterogeneous, diverse audiences (Björgvinsson et al. 2012; Lukyanenko et al. 2016). Second, the field of psychology has investigated the use of conceptual structures that are maximally accessible to non-expert audiences. These are known as *basic level categories* – units of thought and speech that humans find most natural for communication and explanation (Rosch et al. 1976).

Basic level categories are frequently dubbed as among the most important developments in modern psychology (Murphy 2004). Typically, these categories are in the middle of a knowledge taxonomy and represent objects, actions and events. Examples include bird, tree, child, car, cup, table, fish, dress, customer, plane, shopping, watching (TV), swimming, and dancing. The category of "dancing" is less specific than a particular type of dance, such as tango or waltz, but is more specific than the general notion of an "activity". Basic level categories are often an "entry category" – the first thing that comes to mind when encountering a member of the category, or thinking about common objects in a domain (Jolicoeur et al. 1984). They are also commonly among the first words taught by parents, and the first concepts understood

by children (Mervis et al. 1994). It is also notable that more general levels of categories (e.g., animal, furniture) may be less readily accessible to nonexperts, as nonexperts, for example, may struggle to determine if a mushroom is a plant or fungi, while finding few problems recognizing something as a mushroom.

Recently, the benefits of basic level categories have been suggested in the context of IS, in particular, for eliciting requirements and creating appropriate conceptual structures (e.g., database tables, interface design options) for broad and heterogeneous audiences (Castellanos et al. 2020; Lukyanenko et al. 2019). In particular, Castellanos et al. (2020) provided the principles for identifying basic level categories in a domain, while Lukyanenko et al. (2014, 2019) empirically demonstrated the benefits of using basic level categories for designing IS for wide nonexpert audiences.

Toward Basic Explainable AI

We propose that basic level categories hold a unique potential for supporting explainability in AI when explanations are provided for heterogenous, nonexpert audiences. Research in psychology suggests that basic level categories do not presuppose domain expertise. Thus, explanations made using basic level categories are appropriate for general audiences. Indeed, since even children should be familiar with basic level categories, this opens the possibility of extending XAI applications to them as well (e.g., a child asking why a driverless car, or another autonomous agent made a particular decision). Furthermore, as basic level categories are cognitively easy to assimilate (e.g., suggested by the entry category effect), explanations which use basic level categories should also be understood quickly and efficiently.

A simple example of an explanation provided with the use of basic level categories can be in the form of: AI "recommends to sell the house because prices may increase next year", where basic level categories are underlined and determined following the procedure suggested by Castellanos et al. (2020). In contrast, a similar explanation without basic level categories might be in the form of: "valuation of real-estate is expected to soar". The basic premise is the first explanation is more accessible to broader audiences than the second one (we will return to the issue of the semantic differences between the two below).

As virtually all existing approaches to XAI utilize concepts (even when graphical representations are used), basic level categories can be incorporated in all current approaches to XAI. Hence, we adapt the notion of basic level categories to the taxonomy of XAI reviewed earlier and propose a research agenda for future studies.

Research Direction 1: Model-specific Basic XAI. While there is generally a trade-off between explainability and predictive accuracy of AI models, where explainability is of utmost concern, modifications to existing AI routines have been suggested (Adadi and Berrada 2018). Therefore, we suggest investigating approaches for incorporating basic level categories into specific data preparation and training algorithms. One potential solution may be based on performing a dimensionality reduction by aggregating specific information into a common basic level category variable. For example, consider a natural science application of animal observation by citizens, with specific animal information being recorded. Such application may group animals under a basic level category (e.g., specific birds grouped into bird category, specific spiders into a common spider category). This type of aggregation could benefit ML algorithms which do not have their own feature selection capabilities, such as neural networks. Research should also investigate the boundary conditions for such solutions, to ensure that the predictive power of such models remains at an acceptable level for viable applications. Finally, future research is needed on improving the predictive accuracy of such explainability-motivated models.

Research Direction 2: Model-agnostic Basic XAI. This direction perhaps offers the broadest research avenues stemming from the notion of basic level categories. In this direction, the explanations are being generated post hoc, without altering a typically high-performance AI model. Future research can investigate how basic level categories can be employed to enhance the explainability of these existing methods. For example, returning to the Superimposition method (Lukyanenko et al. 2020), future research can investigate the added benefits of simplifying conceptual modeling diagrams by expressing their entity types using basic level categories. For example, if a conceptual model has two entity types (e.g., merchandise manager and logistics manager), a Basic XAI approach would suggest grouping these two entity types under a common entity type "manager". This approach has already been evaluated for increased comprehension in the context of IS interface design (Castellanos et al. 2020). Likewise, the basic level category notion can

be used to enhance the natural language-based approaches to explainability, such as in Krening et al. (2016), whereby scripts in natural languages can utilize as many basic level category terms as possible.

Research Direction 3: Global vs. Local Basic XAI. We expect that an immediate benefit from basic level categories should be accrued when dealing with global, rather than local explanations. In particular, basic level categories can provide a high-level overview of an AI model, abstracting away specific nuances and complexities. Hence, basic level categories can enhance existing methods focusing on global explainability (Adadi and Berrada 2018; Rai 2020).

We expect significantly more challenges when trying to adopt the Basic XAI concept to local explanations. Local explanations typically deal with specific scenarios. Of course, these scenarios can be expressed using basic level category terms, however, more intricate domain specific terminology may be needed. Basic level categories may be too simplistic for many such explanations.

An important direction for future studies under this direction deals with investigations of any negative consequences due to the introduction of basic level categories. Although basic level categories are expected to improve the accessibility of XAI, they invariably simplify the explanation, which in many scenarios can be undesirable. Here, open questions involve the boundary conditions on the use of basic level categories.

Research Direction 4: Challenges in Selecting and Using Basic Level Categories. The final research opportunity relates to the challenges related to the basic level categories themselves. First, determining categories appropriate for XAI contexts (e.g., multi-cultural scenarios) poses challenges and warrants dedicated research. Second, precision is often altered by using basic level categories. For example, precision is lost when replacing a category of "blue footed booby" with the basic term "bird". Likewise, to go from a general category "animal", to the basic level "bird", additional precision needs to be introduced. Research in IS has begun investigating this issue in the context of database and interface design (Castellanos et al. 2020), and can be a source of solutions for Basic XAI.

One open question is whether additional design interventions may help in using basic level categories. For example, explanations can be layered with details gradually revealed to the user upon request, or, even automatically, based on an adaptive process capable of understanding the current comprehension level of the user. In this scenario, basic level categories can reside on an intermediate layer, and serve as a transition point from higher order to more specific explanations.

Conclusion

We propose Basic Explainable AI as an exciting research agenda to make Explainable AI more accessible to nonexpert consumers of AI. Our work carries important implications for research and practice. As the need to explain AI become more urgent, Basic XAI can become a basic mechanism for improving accessibility of these explanations. This should facilitate broader societal adoption of AI. The work further contributes to the theory of basic level categories, extending the boundaries of the theory into the new context of XAI. We encourage the IS community to join us in the development of Basic XAI.

REFERENCES

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. 2018. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An Hci Research Agenda," in CHI Conference on Human Factors in Computing Systems, pp. 1–18.

Adadi, A., and Berrada, M. 2018. "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access (6), IEEE, pp. 52138–52160.

Anderson, J., Rainie, L., and Luchsinger, A. 2018. "Artificial Intelligence and the Future of Humans," Pew Research Center (10), p. 12.

Bailetti, T., Gad, M., and Shah, A. 2016. "Intrusion Learning: An Overview of an Emergent Discipline," Technology Innovation Management Review (6:2).

Björgvinsson, E., Ehn, P., and Hillgren, P.-A. 2012. "Agonistic Participatory Design: Working with Marginalised Social Movements," CoDesign (8:2–3), pp. 127–144.

Castellanos, A., Tremblay, M., Lukyanenko, R., and Samuel, B. 2020. "Basic Classes in Conceptual Modeling: Theory and Practical Guidelines," Journal of the Association for Information Systems (21:4), pp. 1001–1044.

Castelvecchi, D. 2016. "Can We Open the Black Box of AI?," Nature News (538:7623), p. 20.

Crevier, D. 1993. Ai: The Tumultuous History of the Search for Artificial Intelligence, (First Edition.), Basic Books.

Dosilović, F. K., Brčić, M., and Hlupić, N. 2018. "Explainable Artificial Intelligence: A Survey," in International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, pp. 0210–0215.

Gunning, D. 2016. "Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency," Defense Advanced Research Projects Agency (DARPA), nd Web, 2.

Holzinger, A., Kieseberg, P., Weippl, E., and Tjoa, A. M. 2018. "Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable Ai," in International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, pp. 1–8.

Jolicoeur, P., Gluck, M. A., and Kosslyn, S. M. 1984. "Pictures and Names: Making the Connection," Cognitive Psychology (16:2), pp. 243–275.

Krening, S., Harrison, B., Feigh, K. M., Isbell, C. L., Riedl, M., and Thomaz, A. 2016. "Learning from Explanations Using Sentiment and Advice in RL," IEEE Transactions on Cognitive and Developmental Systems (9:1), IEEE, pp. 44–55.

Lukyanenko, R., Castellanos, A., Storey, V. C., Castillo, A., Tremblay, M. C., and Parsons, J. 2020. "Superimposition: Augmenting Machine Learning Outputs with Conceptual Models for Explainable AI," in 1st International Workshop on Conceptual Modeling Meets Artificial Intelligence and Data-Driven Decision Making, Vienna, Austria: Springer, pp. 1–12.

Lukyanenko, R., Parsons, J., and Wiersma, Y. 2014. "The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content," Information Systems Research (25:4), pp. 669–689.

Lukyanenko, R., Parsons, J., Wiersma, Y., and Maddah, M. 2019. "Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content," MIS Quarterly (43:2), pp. 634–647.

Lukyanenko, R., Parsons, J., Wiersma, Y., Sieber, R., and Maddah, M. 2016. "Participatory Design for User-Generated Content: Understanding the Challenges and Moving Forward," Scandinavian Journal of Information Systems (28:1), pp. 37–70.

McCorduck, P. 2004. Machines Who Think, Natick, MA: A. K. Peters, Ltd.

Mervis, C. B., Johnson, K. E., and Mervis, C. A. 1994. "Acquisition of Subordinate Categories by 3-Year-Olds: The Roles of Attribute Salience, Linguistic Input, and Child Characteristics," Cognitive Development (9:2), pp. 211–234.

Miller, T., Howe, P., and Sonenberg, L. 2017. "Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences," ArXiv Preprint ArXiv:1712.00547.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. 2019. "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI," ArXiv Preprint ArXiv:1902.01876.

Murphy, G. 2004. The Big Book of Concepts, Cambridge, MA: MIT Press.

Rai, A. 2020. "Explainable AI: From Black Box to Glass Box," Journal of the Academy of Marketing Science (48:1), Springer, pp. 137–141.

Rao, A. S., and Verweij, G. 2017. "Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise," PwC Publication, PwC, pp. 1–30.

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. Why Should i Trust You?: Explaining the Predictions of Any Classifier, presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp. 1135–1144.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyesbraem, P. 1976. "Basic Objects in Natural Categories," Cognitive Psychology (8:3), pp. 382–439.

Sun, T. Q., and Medaglia, R. 2019. "Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare," Government Information Quarterly (36:2), pp. 368–383.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., and Youngblood, G. M. 2018. "Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation," in 2018 IEEE Conference on Computational Intelligence and Games (CIG), IEEE, pp. 1–8.