LEx: A Framework for Operationalising Layers of AI Explanations

All authors contributed equally to this research.

Ronal Singh
Marc Cheong
Tim Miller
School of Computing and Information Systems
The University of Melbourne
singhrr@unimelb.edu.au,
marc.cheong@unimelb.edu.au,
tmiller@unimelb.edu.au

Upol Ehsan
Mark O. Riedl
Georgia Institute of Technology
Atlanta, Georgia
ehsanu@gatech.edu,
riedl@cc.gatech.edu

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

Several social factors impact how people respond to AI explanations used to justify AI decisions affecting them personally. In this position paper, we define a framework called the *layers of explanation* (LEx), a lens through which we can assess the appropriateness of different types of explanations. The framework uses the notions of *sensitivity* (emotional responsiveness) of features and the level of *stakes* (decision's consequence) in a domain to determine whether different types of explanations are *appropriate* in a given context. We demonstrate how to use the framework to assess the appropriateness of different types of explanations in different domains.

Introduction

The discourse on explainability in algorithmic decision-making has been gaining traction in the past few years. Research in this area proposes various types of explanation methods, such as feature-based explanations, natural language rationales [10], counterfactual & contrastive explanations [16, 4, 22], and directive explanations [23]. Despite the recent progress in Explainable AI (XAI), there is an emerging body of work [2, 20, 24, 9, 14, 9] that show evidence that explanation systems can fail when they do not consider human factors. If the field of XAI is to address the infamous reputation of "inmates running the asylum" [17], where XAI researchers often develop explanations based

on their own intuition rather than the situated needs of their intended audience, we need to go beyond the bounds of the algorithm and incorporate human factors in XAI design [8, 7]. While "opening" the proverbial black-box of AI matters, *who* opens it and *how* the explanations are given matters just as much, if not more than just the *what*.

Recent studies [3, 23] show that certain types of explanations are considered to be *inappropriate* to explain certain aspects of decisions. For example, Singh et al [23] show that for explaining credit-scoring decisions, giving a *directive* explanation that an applicant should find a higher-paying job to increase their income (a directive explanation) is 'condescending' and 'impolite', compared to simply noting that the loan would have been granted *if* the applicant had a higher income (counterfactual explanation). On the other hand, directive explanations were preferred for many other aspects of credit scoring, such as how to optimise the number of credit cards.

Inspired by Nissenbaum's notion of *contextual integrity* [19], which argues from a view that privacy protection should be tied to norms within specific contexts, we propose that explanations in AI should be contextualised as well. We propose a framework called the *layers of explanation* (LEx) framework, which operationalizes the *who, why* and more importantly the *how* of explanations from an AI system. In particular, we can use it as an analytic lens to scope the appropriateness of different types of explanations. Inspired by the results of [3, 23], the framework uses the notions of *sensitivity* of features and the level of *stakes* in a domain to determine whether types of explanations are *appropriate* in a given context. We provide examples of how we can apply this framework to assess explanation appropriateness.

Framework: Layers of Explanation (LEx)

This section defines the rationale behind the LEx framework. The framework proposes three main questions:

1. *'Who?'*: Who are the human agents that receive explanations in the domain?
2. *'Why?'*: Why is the explanation needed and what are the explainability goals?
3. *'How'*: How should explanations be given to specific target segments given their explainability goals?

These questions are framed through the lens of how high the stakes are for a given person in a given context, and how sensitive are the features being referred to.

Stakes and Sensitivity

The two concepts used to determine the appropriateness of the explanations are the stakes and sensitivity. The *stakes* are the consequences (positive or negative, that a person receives for obtaining a particular decision, and its impact on their human agency. We divide the stakes into *low*, *medium*, and *high*. A low stakes domain for a person could be creating a personal account on a social media website, while a high stakes domain is applying for a business loan that can make or break one's financial future.

Sensitivity is the emotional responsiveness or susceptibility of a person to a particular explanation (or feature of an explanation). We divide sensitivity into *low* and *high*. An example of a low sensitivity feature is informing someone that they cannot purchase a ticket because an event is sold out. An example of a high sensitivity feature is referring to someone's ethnic background, particularly for people in a minority group.

Table 2: Defining the layers

[Baseline] No explanation (decision only): The automated systems communicates (only) the decision.

[Layer 1] Feature-based explanation: The explanation states the features relevant to a decision.

[Layer 2] Contrastive explanation: The explanation provides not only the features and values but also states how the values of the features need to be different for a different outcome or decision.

[Layer 3] Directive explanation: The explanation lists all of the information from previous layers and suggests *actions* or *interventions* that could bring about the desired outcome or decision.

Note that sensitivity is not the same as a *protected feature*[1], which is an feature that is (often legally) prevented from being used as part of a decision. A sensitive feature may be legal, but inappropriate. Crenshaw's Intersectionality framework [6] may also be relevant from the viewpoint that features interact or overlap with each other. The intersecting combinations of features may directly shape one's circumstance and plays a key role in compounding the sensitivity of sets of feature(s).

Consider the example in the previous section: a sensitive feature may be high for one group, but low for another; ethnicity is a sensitive feature, but may be less sensitive in some domains/contexts for someone with a dominant culture. Hence, designers of XAI systems, especially of the dominant culture and privileged epistemic positions [12] in society, need to suspend all preconceived knowledge of their own socio-historical and personal circumstances. Table 1 summarises the six potential combinations of sensitivity and stake, along with illustrative examples.

Defining the layers

We propose the *layers of explanation* (see Table 2), inspired by various types of explanations [13, 1, 15, 18], including Singh et al.'s [23] notion of directive explanation. We also provide examples showing how to develop the layers and assess the appropriateness given prior knowledge of how people may respond to these explanations.

Note the ordinal relationship between layers: if one layer is inappropriate or sensitive, any higher layer will also be; e.g. directive explanations refer to counterfactuals, so if a counterfactual explanation is considered insensitive, then so too will the directive explanation.

[1] See e.g. definitions in the Australian legal context: https://www.fairwork.gov.au/employee-entitlements/protections-at-work/protection-from-discrimination-at-work

Table 1: Stake versus feature sensitivity: examples

	Feature Sensitivity	
Stake	Lower	Higher
Low	Social media account registration: A user cannot create a new account due to anti-spam heuristics; the user is asked to retry.	Recommender systems: New products are recommended based on a user's purchase history/demographics.
Medium	Exam grading: A student's writing score in an exam is auto-graded based on historical patterns.	Automated Recruitment: Job outcome for an applicant is due to a protected feature.
High	Loan: A small trader's loan application has been denied due to certain business rule.	Bail: An accused person is denied bail (e.g. due to statistical correlation to a protected feature).

Process

In this section, we discuss the process of the LEx framework to judge the appropriateness of explanations.

The process model is straightforward. First, we answer the *'why?'* and *'who?'* for the domain to identify the list of potential explainees and the goals of explanation for each of these. The layers are used to answer the *'how?'*, where the answer to these involves identifying which layer

*To ensure that the outcome of these questions remains impartial, we recommend using the idea of the *Original Position* (OP), proposed by political philosopher John Rawls [21]. The "most appropriate moral conception of justice" [11] is obtained when the parties take up the "*veil of ignorance*", completely depriving themselves of all knowledge of their own personal circumstances and attributes; in short, putting themselves in the shoes of others.

is appropriate*.

High stakes example 1: AI-based credit scoring
The first example is for a typical algorithmic decision-making case study [5]. Consider a case in which the 'who?' question is an applicant whose loan application has been rejected, and the 'why?' is for them to learn how they can get approval in future. The different layers are:

Baseline No explanation: 'Your loan has been denied'.

Layer 1 Feature-based: 'The decision was made based on these variables: income, with these weights: . . . '.

Layer 2 Contrastive: 'The loan may be approved if the applicant were to have an income of more than . . . '.

Layer 3 Directive: 'The loan may be approved if the applicant were to have an income of more than . . . . The applicant could get a second job or ask for a promotion to increase the income.'

Credit risk assessment is a high stakes domain. We know from earlier works [3, 23] some explanations are deemed inappropriate. While *income* is a legitimate feature for the AI to use, its use can be offensive depending on *how* it is used in an explanation. As such, we may decide to provide explanations up to Layer 2 (contrastive). However, explaining is a social or interactive process [17], and we could increase the *level of verbosity* (go to Layer 3) if the recipient requests further details.

Low stakes example: IT Admin assisting an employee with a disabled user account
In the following example, the domain and feature are both in the lower quadrants; we could offer Layer 3 explanation.

Layer 1 'Account disabled due to login attempts.'

Layer 2 'Account disabled due to more than 10 login attempts.'

Layer 3 'Account disabled due to more than 10 login attempts. To resolve, either reset the login attempts counter or wait for 48 hours for an automatic reset.'

Final Remarks

While *who* we explain to is critical [9], *how* we explain is equally important. Studies [3, 23] reveal that current XAI tools present inappropriate explanations to people when using certain features. Building on the *layers of explanation* (LEx) framework that enables practitioners to judge an explanation's appropriateness, we need to further explore the following aspects:

1. What factors make an attribute sensitive for an individual, and how should we operationalise these factors in an explanation generation pipeline to align with one's explanatory needs?
2. How do these sensitivity factors change with application domain?
3. What design guidelines should we incorporate during model development and deployment pipeline to empower users with the agency to decide the *what, when,* and *how* of an explanation?

By tackling these questions, we can refine the bounds of explainability in a human-centered manner, one that accommodate societal expectations and cater to the human factors that govern people's reaction to AI-mediated explanations.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[4] Ruth MJ Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning.. In *IJCAI*. 6276–6282.

[5] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An Interpretable Model with Globally Consistent Explanations for Credit Risk. (Nov. 2018). http://arxiv.org/abs/1811.12615

[6] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Univ. Chic. Leg. Forum* (1989), 139. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/uchclf1989&section=10

[7] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

[8] Upol Ehsan and Mark O Riedl. 2019. On Design and Evaluation of Human-centered Explainable AI systems. In *Proceedings of Emerging Perspectives in Human-Centered Machine Learning : A Workshop at The ACM CHI Conference on Human Factors in Computing Systems*.

[9] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. *arXiv preprint arXiv:2002.01092* (2020).

[10] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *Proceedings of the International Conference on Intelligence User Interfaces*.

[11] Samuel Freeman. 2019. *Original Position* (summer 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/original-position/

[12] Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Clarendon Press.

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2019), 93.

[14] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for

Explainable AI User Experiences. *arXiv preprint arXiv:2001.02478* (2020).

[15] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[16] Tim Miller. 2018. Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163* (2018).

[17] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[18] Christoph Molnar and others. 2019. Interpretable machine learning: A guide for making black box models explainable. *E-book at< https://christophm. github. io/interpretable-ml-book/>, version dated* 08 (2019).

[19] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review* 79 (2004), 119.

[20] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[21] John Rawls. 1971. *A Theory of Justice*. Harvard University Press.

[22] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.

[23] Ronal Singh, Paul Dourish, Piers Howe, Tim Miller, Liz Sonenberg, Eduardo Velloso, and Frank Vetere. 2021. Directive Explanations for Actionable Explainability in Machine Learning Applications. (2021).

[24] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.