Pragmatic AI Explanations

Shi Feng, Chenhao Tan
University of Chicago
{shif, chenhao}@uchicago.edu

1 AI should Explain Pragmatically

To explain is to help the *listener* understand. A pragmatist would interpret explanation as a means to strengthen the listener's understanding of the speaker's worldview [9, 16, 21, 23]. Despite certain oversimplifications, this pragmatic view emphasizes an important aspect of effective explanations: to be compatible with the listener's existing worldview and to provide digestible information. This paper examines AI explanations from a pragmatic perspective and argue for explicit treatment of the listener in the generation of explanations. In short, we propose that AI should explain pragmatically.

As an motivating example, consider using explanation as a recourse for a loan applicant who got rejected by an AI system. What the applicant seeks in recourse is not so much why they are rejected, but what they can do differently to get an approval. So using input attribution methods [13, 19] to list features of the applicant as evidence for rejection is not effective, as it does not provide actionable feedback. Instead, "increase annual income by \$5k" or "reduce risky assets in portfolio" can be good explanations, while "decrease age by 2" or "increase income ten fold" are not. Discerning good explanations from bad ones requires knowing what the applicant can plausibly achieve, e.g., knowing that increasing income by certain amount is feasible for some but not others. Whereas loan approval is about predicting a single aspect about the applicant, i.e., the likelihood of repayment, it takes a more comprehensive understanding of the applicant to generate a good explanation. So how can we characterize the knowledge gap between these two problems? And what kind of data can inform the AI system to generate more actionable explanations?

We build on the Rational Speech Act framework (RSA, [7]) and the level-k model of reasoning in Keynesian game theory [17, 20]. We argue that existing AI explanations are not pragmatic, and identify types of data that can inform AI pragmatic reasoning. Previous work on explanatory dialogues [1, 11] also took inspiration from pragmatics literature since the need of modeling the listener arises naturally from the dialogue format. In contrast, we study individual explanations independent of their potential dialogue context. Our paper provides theoretical grounding for existing work on adaptive explanations [1] and presents two concrete research projects on incorporating pragmatics into XAI [15].

2 Existing AI Explanations are not Pragmatic

We use RSA to model AI explanations. RSA starts with a level-0 listener L_0 who interprets a message x *literally* according to a world model $\mathcal{L}$. We can define $\mathcal{L}(x, w) = \mathbb{1}[x]^w \cdot P(w)$, where the indicator function $\mathbb{1}[\cdot]$ is one if the message x evalutes true in world state w and zero if otherwise. Using L_0 as the internal model of listener, we have a level-0 speaker S_0 who simulates L_0 and chooses a message that's maximally useful for L_0. The utility of a message is typically measured by how much it reduces L_0's information-theoretic uncertainty about the world state w: $\ln P_{L_0}(w \mid u)$. A level-1 listener L_1 assumes that the speaker is rational and trying to optimize informativness, so L_1 uses S_0 as the internal model of the speaker, and infers the speaker's intended meaning—the world state being conveyed—following Bayes's rule. Continuing the recursion, a level-1 speaker S_1 then uses L_1 as the internal model and chooses an explanation that's maximally useful for L_1. The whole

Preprint. Under review.

inference is conditioned on the w which is the world state according to the speaker.

$$P_{L_0}(w \mid x) \propto \mathcal{L}(x, w) \qquad P_{S_0}(x \mid w) \propto \exp\{\alpha_{S_0} \cdot \ln P_{L_0}(w \mid x)\}$$
$$P_{L_1}(w \mid x) \propto P_{S_0}(x \mid w)P(w) \qquad P_{S_1}(x \mid w) \propto \exp\{\alpha_{S_1} \cdot \ln P_{L_1}(w \mid x)\}$$

Pragmatic reasoning emerges at level-1 for both listener and speaker. The internal simulation of level-1 pragmatic listeners allows them to draw inferences about a whole host of implicated meaning beyond the literal meaning of a linguistic expression. For example, the listener can infer the speaker's intention, e.g., whether they are explaining objectively or trying to persuade. The inference leads to biases in how the listener interprets the speaker: the same explanation can be perceived as trustworthy or deceptive depending on the listener. The recursive self-reference in the level-1 pragmatic speaker's simulation (through the internal model based on L_1) allows them to reason about and properly respond to these biases of the listener. Level-1 speakers are *self-aware* in the sense that they understand the listener's biases about them and adjusts their explanations accordingly.

The lack of such self-awareness dictates that existing AI explanations do not qualify as level-1 speakers; they are not pragmatic. In a loose sense, the prediction model can be thought of as approximating an aspect of the world model $\mathcal{L}$ by training on human labeled dataset, and the explanation method resembles the level-0 speaker. Note that our categorization is a statement about the speaker's mechanics, not their product. In some cases, level-0 speaker might provide an explanation that's also optimal at level-1.

The main caveat of this framework arises from RSA's oversimplifications. RSA was originally introduced in the context of referential game which has many simplifying assumptions that don't directly translate to the scenarios of AI explaining to humans. RSA typically treats $\mathcal{L}$ as a common grounding and assumes it is known to both parties, which is a reasonable assumption in referential games. But this assumption is not generally true due to each person's private knowledge and personal beliefs. And we cannot claim with certainty that any such common grounding exists between human and AI. So the $\mathcal{L}$ in the AI's pragmatic reasoning is inevtitably an approximation of the human's world model, and the mismatch with the real human $\mathcal{L}$ is hard to quantify. It's debatable, in particular, whether modeling the pragmatic process is necessary as opposed to treating everything as part of $\mathcal{L}$ and approximating it. Approximating the part of $\mathcal{L}$ that determines the interpretability of an explanation can be thought of as learning an interpretability prior [10]. The linguistic community has considered extending RSA with epistemic access [2], but those aren't directly applicable to our scenario. We believe our framework provides novel and important insights despite this caveat. In particular, it offers a new perspective into data collection. As we discuss in the next section, treating AI as a pragmatic speaker reveals fatal flaws in existing schemes of data collection and how explanations are deployed.

3 Building Level-1 AI Explanations

We outline two proposals for building level-1 explanations. For each proposal, we examine an existing method. These methods were not developed under a pragmatic framework, but we treat them as speakers in RSA, and *retrospectively* identify their underlying level-0 assumptions. We discuss how these hardcoded assumptions can be violated in reality, and as a antidote, how they can replaced by learning from two types of level-1 data, one in the form of feedback and another in direct supervision.

3.1 Learning from L_1 feedback with online optimization

The most direct way to build level-1 explainer is to collect level-1 feedback from L_1, i.e., data collected by deploying the explanations and observing how listeners react to them. To motivate this approach, consider LIME [19], a representative method for input attribution, which explains a prediction by highlighting regions of the input, e.g., phrases in a sentence for text classification or regions in an image for object recognition. For each input, LIME assigns a saliency score to each of its units (e.g., a word in a sentence, or an image segment) which measures the importance of that unit for the prediction. Visualizations of LIME scores typically use a fixed colormap to translate real-valued saliency scores to colors, and generates a heatmap over the input. A common choice is a diverging colormap where blue represents positive contribution by the unit to the prediction, red represents negative contribution, and white is neutral or no contribution; darker colors imply stronger contribution [5, 12, 22]. This design choice makes an implicit level-0 assumption about

how the listeners preceive color: that the listenes can distinguish these colors, that they make the right associations (blue with positive, red with negative), and that they correctly associate shades of color with numerical values. Whether these assumptions hold depends on an individual's physical limitation (e.g., colorblindness), cultural background [14], and personal experience [8]. This leads to variance in the efficacy of explanations to different listeners [6]—a variance that, once controlled for, can lead to improvement in overall explanation effectiveness.

Controlling for such variance requires first identifying an action space. Using the colormap example, an action space is a set of options for one aspect of the colormap, such as the choice of two diverging colors, or the mapping between real-valued numbers to the color darkness. Then, we can collect stratified data of listener's response to explanations in a user study, e.g., 50% users see blue-red and 50% see green-red. This type of data allows us to perform online optimization and choose the best colormap for each user and improve explanation quality with more an more interaction with the user.

3.2 Learning from S_1 supervision by thinking in AI's shoes

Another line of work collects human rationale for classification problem, examplified by e-SNLI [3]. Annotators for those datasets are tasked to highlight words in the input that are most important for the label and provide explanations.

But there is an important mismatch that's not hard to identify under the pragmatics lens. In data collection, the human annotators explain as if they are talking to a human listener who thinks the speaker is also human. In RSA terms, the human annotator's pragmatic inference uses an $\mathcal{L}$ which encodes the social norm of human-human conversation. But when AIs explains, the human listener knows that the speaker is AI, and their $\mathcal{L}$ will change accordingly. So by training the AI to mimic explanation generated from the first process, we are putting our AI under a false impression that the listeners will treat the same as a human speaker.

To fix this, we must think like an AI. When providing human explanation data, we must explain as if the listener thinks they are talking to an AI. In RSA terms, we should construct explanations where our approximation of $\mathcal{L}$ encodes the social context of human-AI conversation, rather than human-human conversation. This can be done by providing an interface for human to simulate AI [18]. The data created this way can provide direct S_1 supervision for the AI as an level-1 speaker.

This proposal aligns nicely with the discussion on data granularity in robustness [4]. We ask the question: what data should be treated as level-k? The answer hinges on what we expect a piece of data to generalize: level-0 data is what we expect to generalize for all listeners, while level-1 data is specific to each listener but also each speaker, since the whole inference is dependent on the world state w according the speaker's belief.

3.3 The social explainer

Can AIs generate level-1 explanations without becoming true level-1 progmatic reasoners? We think yes, if the human listener is properly conditioned and the AI follows certain $\mathcal{L}$. A naive solution is to fool the human listener and create and illusion that they are talking to a human instead of an AI, and make sure the AI can maintain such illusion—a capability that's already within the reach of existing AI systems. If we can do this successfully, the issues discussed in the two previous proposals wouldn't be a problem. But once the illusion is broken, the knowledge that we attempted to deceive the human listener can lead to catestrophic outcomes. Perhaps a more honest and safe approach is to inform the AI of its social perceptions, and ask it to explain accordingly.

4 Conclusion

RSA provides a starting point for modeling AI explanation as a pragmatic inference process. Examining AI explanations under the pragmatics lens reveal fatal flaws in how we currently train and deploy AI explainers. To evolve from level-0 to level-1, we present two proposals for data collection and training: learning from L_1 feedback, and learning from S_1 supervision.

References

[1] Arjun R Akula, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Y Chai, and Song-Chun Zhu. X-tom: Explaining with theory-of-mind for gaining justified human trust. *arXiv preprint arXiv:1909.06907*, 2019.

[2] Richard Breheny, Heather J Ferguson, and Napoleon Katsos. Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, 126(3):423–440, 2013.

[3] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *Proceedings of Advances in Neural Information Processing Systems*, 2018.

[4] Chen Cheng, Hilal Asi, and John Duchi. How many labelers do you have? a closer look at gold-standard labels. *arXiv preprint arXiv:2206.12041*, 2022.

[5] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the Association for Computational Linguistics*, 2017.

[6] Shi Feng and Jordan Boyd-Graber. What can AI do for me: Evaluating machine learning interpretations in cooperative play. In *International Conference on Intelligent User Interfaces*, 2019.

[7] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[8] John Gage. *Color and meaning: Art, science, and symbolism*. Univ of California Press, 1999.

[9] Peter Gärdenfors. A pragmatic approach to explanations. *Philosophy of Science*, 47(3):404–423, 1980.

[10] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *arXiv preprint arXiv:1805.11571*, 2018.

[11] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*, 2022.

[12] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

[13] Scott Lundberg and Su-In Lee. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv: 1611.07478*, 2016.

[14] Thomas J Madden, Kelly Hewett, and Martin S Roth. Managing images in different cultures: A cross-national study of color meanings and preferences. *Journal of international marketing*, 8 (4):90–107, 2000.

[15] Tim Miller. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[16] Theodore Mischel. Pragmatic aspects of explanation. *Philosophy of Science*, 33(1/2):40–60, 1966.

[17] Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326, 1995.

[18] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *International Conference on Human Factors in Computing Systems*, 2021.

[19] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining*, 2016.

[20] Dale O Stahl and Paul W Wilson. On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

[21] Bas C Van Fraassen. The pragmatics of explanation. *American philosophical quarterly*, 14(2): 143–150, 1977.

[22] Eric Wallace, Shi Feng, and Jordan Boyd-Graber. Interpreting neural networks with nearest neighbors. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[23] Noga Zaslavsky, Jennifer Hu, and Roger P Levy. A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*, 2020.