Pragmatic Intelligence Augmentation

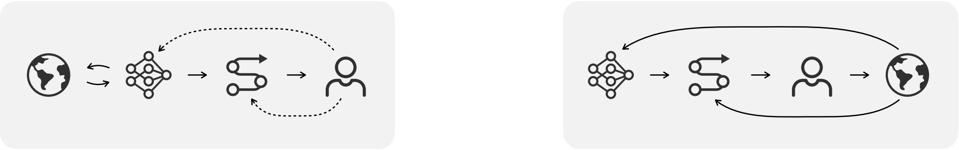
Shi Feng — Research Statement

My overarching goal is intelligence augmentation with machine learning—to enhance and expand human capabilities, make impossible tasks possible, and make difficult tasks easier. I'm particularly interested in tasks where humans play an indispensable role that cannot be automated despite machines becoming times more capable than they are now. I advocate for a paradigm shift: instead of training models to emulate humans and solve tasks autonomously, train models to support humans.

A key challenge of leveraging machine learning is the lack of interpretability: it remains difficult to reliably answer questions about how models come up with their predictions. And this issue is particularly salient in the augmentation paradigm where a semantically meaningful interface to control these models becomes a necessity. Interpretability's role is to define a shared language between human and machine learning models. But what do these supposedly intepretable models really know about humans? How can we inform them about what we want to achieve and what information we need?

The critical missing piece, in my opinion, is *pragmatics*. Consider the process of using a machine learning to support human decisions. The physical barrier between human and machine divides this process into two: *explanation* is the machine process to generate an output, and *interpretation* is the human process of making a decision based on that output. The role of pragmatics is to create a model within the machine to simulate the human interpretation process. Given this model, the machine can explain pragmatically by generating explanations to optimize the objective based on the expected human decision according to that model.

Pragmatic explanation is thus the inverse problem of interpretation. The simulation of human interpretation by a *listener model* shares the same spirit as RSA from linguistics [9] and level-k model from economics [11]. In the context of intelligence augmentation, a useful listener model should learn about the human decision maker's goals and values as well as constraints and limitations, e.g., the budget in terms of time and effort, the human's background knowledge about the task or the lack thereof.



(a) Emulation feedback + pragmatic info hard-coded in one-shot

(b) Pragmatic info learned from feedback

Figure 1: In existing interpretability work (a), the model's output is presented to humans through a post-processing step (e.g., post-hoc explanation methods). The human remains outside the feedback loop, and the only information that supports machine pragmatic inference is injected to the system prior to any interaction with the human, as depicted by the dashed arrows. My proposal (b) puts the human in the feedback loop where their decision given model output is directly evaluated, modeled, and optimized.

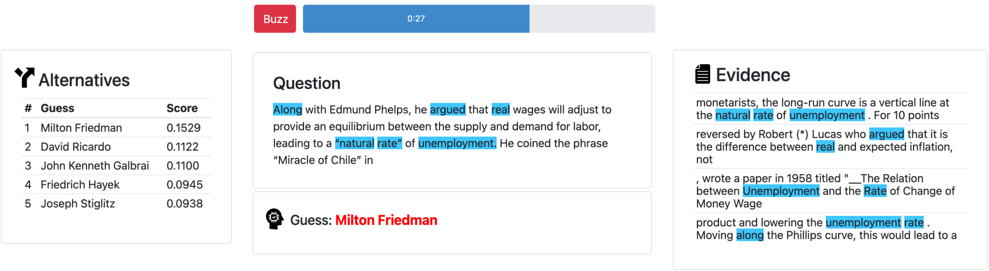
Existing systems have very limited access to this information; they use naive (implicit) listener models built on a set of heuristics. As depicted in Figure 1(a), such heuristics are hard-coded into the system either as post-processing steps (i.e., post-hoc explanation methods) or as regularizations (i.e., inherently interpretable architectures). The only information that can support machine pragmatic inference is all injected into the system prior to any interaction with humans; the system has no mechanism to learn from interaction.

My approach: instead of hard-coding pragmatic knowledge into the listener model, build mechanisms with which the listener model can be learned; instead of building models that are useful, build ones that *aspire* to be more useful to humans over time. I have focused on two main lines of research. **R1** focuses on closing up and learning from the pragmatic feedback loop [5, 6]. In particular, I demonstrate how existing explanation methods—albeit flawed—can serve as building blocks for a human–AI team that learns to outperform both human experts and autonomous AIs. **R2** focuses on preemptively identifying pitfalls in human–AI cooperation and improving the worst-case performance of the team. I discover pathologies of machine learning models that make cooperation difficult [8, 7, 16, 18], build mechanisms to expose them through interaction with humans [17], and mitigate them [8, 15, 14, 21, 20].

R1. Learning in a pragmatic feedback loop

Figure 1(b) provides a general scheme for learning to be more useful for the human. But to close the feedback loop we must specify mechanisms for collecting and integrating feedback. **Collecting feedback** is similar to an application-grounded evaluation [3] of the explanation. But an additional challenge is that the human must be incentivized to actively engage with the system—if they pay little attention to the system, their feedback won't help the system. **Integrating feedback** requires a mechanism to store and update information collected from interaction. I take a model-based approach and use an explicit listener model to parameterize such knowledge. Instead of a cold-start, we can initialize the listener model using the bag of heuristics employed by existing explanation methods, and adapt it for each individual through interactions.

Learning from feedback requires trial-and-error. But these might be an opportunity cost associated with providing suboptimal support. For critical problems such as cancer diagnosis, no clinical trial would be approved if there is a risk of lower quality care for the patient. For less high-stake scenarios where we can afford to trial-and-error, it is still crucial to model the **exploration–exploitation** trade-off.



Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

Guess: Milton Friedman

Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment . For 10 points

reversed by Robert (*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "__The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

Figure 2: Our web interface for human-AI Quizbowl. The question is displayed in the middle; the list of alternative answers on the left provides additional detail about the model's confidence estimate. example-based attribution on the right shows relevant training examples; words in both the question and evidence are highlighted by input attribution. Each explanation can be displayed or hidden individually.

In my two-part work, `Human-AI Quizbowl` [5, 6], I use a trivia game as testbed and finish a complete setup of closing up and learning from the feedback loop. For all my experiments, I conduct user studies with expert human players competing against each other, just like they do in real-life games. The first part focuses on collectin feedback, and the second part focuses on integrating feedback and closing the loop.

In the first part [5], I build a web interface (Figure 2) for Quizbowl, a trivia game where players compete to see who can answer a question correctly with the least amount of information. Unlike in traditional Quizbowl where each team is made of one or more human players, each team in our game is a single human paired with a machine learning system. The strengths of the system can complement the human: it's not as good as humans at connecting the dots, but it's excellent at memorizing facts. The human players are incentivized to engage with the system because they may gain a competitive edge in the game.

With randomized trials where players see random combinations of explanations, we validate that explanations indeed improve human decision quality: players with assistance score higher than those without. But the benefit of explanations is not universal: the team performance shows high variance for different explanations and different questions. Importantly, experts generally benefit from seeing multiple types of explanations at the same time, but novices can be overwhelmed by too much information and fail to reject incorrect model predictions. The explanation's effect depends on who is answering which question, so applying the same set of explanation methods for all decisions is suboptimal. This gives us an idea for how to integrate the feedback into the system: learn to explain *selectively.*

In the second part, `Selective-explanations` [6], we build a listener model that predicts the score a player gets on a question given certain model explanation. Before the start of each question, the listener makes this prediction for each possible configuration of explanations available on our interface, and each configuration receives a score. A selector then picks the best configuration based on the scores and the interface is arranged accordingly. The selector can choose to show nothing if, for example, the speaker model predicts that explanations can mislead the player. Once the player provides an answer, we observe an immediate feedback, and the listener model updates accordingly. Initially the system knows nothing about its human teammate, so it will likely provide suboptimal support, but the hope is that it improves over time. To model the exploration-exploitation trade-off, we formulate this online optimization problem as contextual multi-armed bandit, and use the total score from a game with multiple questions as the objective. Experiments with expert players validate that selective explanations indeed outperform static explanations as well as human player without assistance and autonomous AI players.

R2. Improving the worst-case in human-AI cooperation

The emergence of various cognitive functions in machine learning systems makes it tempting to anthropomorphize them—that the system can *learn* from a few demonstration examples [1], and that they can *explain* their predictions [13]. But labeling system behavior with often not so well-defined terms can create blindspots in human-AI cooperation: when the system stops behaving intuitively, the human's decision becomes difficult: is this a bug, is there artifacts in the data that the model is exploiting, or did the model discover some genuine new insight, like AlphaGo's move 37? To make these decisions easier and preempt fallouts, it's important to understand the degree to which models truly resemble human reasoning.

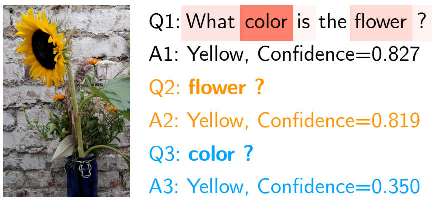

Q1: What color is the flower ?
A1: Yellow, Confidence=0.827
Q2: flower ?
A2: Yellow, Confidence=0.819
Q3: color ?
A3: Yellow, Confidence=0.350

Figure 3: A VQA model makes a correct prediction on the original question Q1. Input reduction finds to Q2; the original prediction no longer makes sense, but the model remains highly confident. Comparing Q2 and Q3, we see a behavior that's inconsistent with the explanation for Q1 which suggests that "color" is more important than "flower".

Under this theme, my first sequence of work starts by examining the central claim of input attribution methods: that they can highlight words *important* to the model's prediction. I ask: important in what sense? Is the definition of importance by perturbation-based [13] and gradient-based methods [4] consistent with our intuition? Importance can be thought of as necessity and sufficiency: a word is important if altering other words has little effect on the model. In `Pathology` [8], I take this intuition to test by iteratively removing words with low attributed importance from the input until the model prediction changes; intuitively, the important ones should remain at the end. But to our surprise (Figure 2), words with highest attributed importance are gone and what's left behind is nonsensical gibberish, yet the model remains confident.

This counterintuitive phenomenon has two root causes. First, attributed importance is derived from model confidence on counterfactual inputs. But the confidence is often ill-calibrated especially on out-of-distribution (OOD) data points. And removing a supposedly unimportant word can often lead to OOD inputs. So the explanations become inconsistent with model prediction on counterfactuals. Secondly, the attribution of each word is computed individually so it ignores higher order interactions between words. So, for example, the effect of altering a phrase in the input cannot be reliably predicted from the attribution of individual of words. The root causes are largely independent of the system's regular accuracy. Although we only evaluated LSTMs (the paper was written before ELMo [12]), the issue persist in the latest Transformer models as validated by independent reproduction of our study on many more tasks [10, 19].

The root causes motivated several mitigation methods. My first fix proposed in `Pathology` [8] to improve calibration on OOD data by a regularization objective. Attribution benefits from repressing catastrophically bad confidence estimates on counterfactual inputs. `Conformity-LIME` [15] replaces softmax score with conformity score for confidence estimation in perturbation-based attribution. `CASO` [14] produces context-aware attribution by considering the second-order interaction between input units. And finally in `Trick-me-if-you-can` [17], we demonstrate how attribution in an interactive interface can help humans uncover model shortcuts and compose stress-test examples to validate those shortcuts.

My second sequence of work examines the claim that large language models can *learn* from context. In particular, I study the models' hyper-sensitivity to certain cues. In `Triggers` [16], we discover special token sequences that cause language models to ignore the context and produce an output distribution that's extremely skewed. For example, GPT-2/3 models become highly offensive when triggered by "TH PEOPLEMan goddreams". Triggers automatically discovered by our method are primarily short nonsensical gibberish token sequences, and the model's response is in stark contrast to how little attention humans would pay to them. This work is done before GPT-3 and the rise of in-context learning or prompting, but our results generalize well. More importantly, this hyper-sensitivity is concerning if we want the model to reliably learn from a few examples like humans do. And indeed GPT-3 is sensitivity to nonessential properties of the demonstration, such as the ordering of examples. In `Calibrate-before-use` [21], we introduce a highly effective heuristic to calibrate and stabilize the model without any additional labeled data. The method is very similar in spirit to the regularization method from `Pathology`: it calibrates the model on inputs that carry no information about the prediction. In a sense, we are teaching the model to ignore non-predictive cues. In `Active-example-selection` [20], we train example selection policies with reinforcement learning to further stabilize in-context learning from demonstrations.

Other future work

In addition to extending the above mentioned two lines of work (see the discussion section of each paper for detail), I'm also actively working on the following open problems in pragmatic intelligence augmentation:

- `Learn-the-prior`: To learn an initialization for the listener model rather than a bag of heuristics, e.g., by learning people's general preference for the length of explanation for different domains. Preliminary experiments have validated this idea on synthetic tasks [2] and a series of machine-assisted classification problems with crowd workers, including pneumonia diagnosis from chest x-ray.
- `Learning-to-teach`: I study a model-based approach for improving flashcard learning. Similar to how we initialize the listener model from a bag of heuristics in `human-AI Quizbowl`, here we initialize a learner model from spaced repetition heuristics. We then update the learner model through recommending flashcards to the learner and collecting feedback.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] C. Chen, S. Feng, A. Sharma, and C. Tan. Machine explanations and human understanding. *arXiv preprint arXiv:2202.04092*, 2022.

[3] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *Springer Series on Challenges in Machine Learning*, 2018.

[4] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the Association for Computational Linguistics*, 2017.

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

[6] S. Feng and J. Boyd-Graber. Learning to explain selectively: A case study on question answering. In *Proceedings of Empirical Methods in Natural Language Processing*, 2022.

[7] S. Feng, E. Wallace, and J. Boyd-Graber. Misleading failures of partial-input baselines. In *Proceedings of the Association for Computational Linguistics*, 2019.

[8] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of Empirical Methods in Natural Language Processing*, 2018.

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

[10] S. Longpre, Y. Lu, and C. DuBois. On the transferability of minimal prediction preserving inputs in question answering. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

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

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining*, 2016.

[14] S. Singla, E. Wallace, S. Feng, and S. Feizi. Understanding impacts of high-order loss approximations and features in deep learning interpretation. In *Proceedings of the International Conference of Machine Learning*, 2019.

[15] E. Wallace, S. Feng, and J. Boyd-Graber. Interpreting neural networks with nearest neighbors. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[16] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of Empirical Methods in Natural Language Processing*, 2019.

[17] E. Wallace, P. Rodriguez, S. Feng, I. Yamada, and J. Boyd-Graber. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. In *Transactions of the Association for Computational Linguistics*, 2019.

[18] E. Wallace, T. Z. Zhao, S. Feng, and S. Singh. Concealed data poisoning attacks on nlp models. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

[19] P. Zhan, Y. Wu, S. Zhou, Y. Zhang, and L. Wang. Mitigating the inconsistency between word saliency and model confidence with pathological contrastive training. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2226–2244, 2022.

[20] Y. Zhang, S. Feng, and C. Tan. Active example selection for in-context learning. In *Proceedings of Empirical Methods in Natural Language Processing*, 2022.

[21] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference of Machine Learning*, pages 12697–12706. PMLR, 2021.