# From Explainability to Explanation: Using a Dialogue Setting to Elicit Annotations with Justifications

**Nazia Attari**
Research Institute for Cognition and Robotics
Bielefeld University, Germany
nattari@techkfak.uni-bielefeld.de

**Martin Heckmann**
Honda Research Institute Europe
Germany

**David Schlangen**
Computational Linguistics
University of Potsdam, Germany

## Abstract

Despite recent attempts in the field of *explainable AI* to go beyond black box prediction models, typically already the training data for supervised machine learning is collected in a manner that treats the annotator as a "black box", the internal workings of which remains unobserved. We present an annotation method where a task is given to a pair of annotators who collaborate on finding the best response. With this we want to shed light on the questions if the collaboration increases the quality of the responses and if this "thinking together" provides useful information in itself, as it at least partially reveals their reasoning steps. Furthermore, we expect that this setting puts the focus on *explanation* as a linguistic act, vs. *explainability* as a property of models. In a crowd-sourcing experiment, we investigated three different annotation tasks, each in a collaborative dialogical (two annotators) and monological (one annotator) setting. Our results indicate that our experiment elicits collaboration and that this collaboration increases the response accuracy. We see large differences in the annotators' behavior depending on the task. Similarly, we also observe that the dialog patterns emerging from the collaboration vary significantly with the task.

## 1 Introduction

Imagine asking a friend whether you can borrow their car for the afternoon, and the only reply you get is "no". You would presumably perceive this as somewhat brusque, and Conversation Analysis would back you up there: This kind of *dispreferred reply* typically needs more work, often being initiated with a filled pause, and being augmented with a *reason* for the refusal (Schegloff, 2007; Levinson, 1983). Now imagine you are asking a car rental place, via their website, whether you can rent a car for the afternoon, and again all you get as a reply is a "no". You would still not be pleased, but the difference here would be that while your friend may have been un*willing* to tell you their reasons, the car rental company, having used a complex statistical model that judged you untrustworthy, based on various kinds of information it has about you, would be un*able* to state reasons (other than a vacuous one like "your score is too low").

The field of *explainable AI* has set itself as a goal to open up the blackbox of current prediction models in order to make their decisions more transparent and also identifying problems concerning the core issues in AI safety. (See (Gilpin et al., 2018; Doshi-Velez and Kim, 2017; Ribeiro et al., 2016; Lundberg and Lee, 2017; Amodei et al., 2016) for recent overviews.) The focus there typically is on providing explanations of decisions in terms of examples or secondary models (e.g. (Kim et al., 2018; Letham et al., 2015; Yuan et al., 2019; Zhang et al., 2019)), where the resulting explanations are understandable at best to experts. In contrast, our interest is in learning to provide verbal explanations, accessible also to novice users. As a first step, we are interested in methods for eliciting data that can be used for this. In this paper, we present an annotation scheme where a pair of annotators works in collaboration to find the best response to a question. Our hypothesis is that a) this leads to better quality responses compared to non-collaborative annotation, as the annotators can actively acknowledge/correct/help their partners, b) the resulting discussions give access to the collaborative thinking directions that lead to the final response, and c) puts the focus on *explanation* as a linguistic act, vs. *explainability* as a property of models. We present results from three different annotation tasks. For each task we compare the accuracies of the responses we obtain in a dialog (two annotators) and a monologue (one annotator)

setting, analyze to what extent the task triggered discussions in the dialog setting and quantify dialog patterns emerging in the interaction of the annotators.

## 2 The Annotation Game

We formalise the annotation task as a game with the following structure. The annotation problem is posed by a special participant in the game, which we call *Nature* ($N$). $N$ poses a question $Q$ that is to be answered, and provides relevant information $I = \{i_1, \ldots, i_n\}$. (e.g., $Q$ = "what is in this image?", with $I$ consisting of an image.) Besides $N$, there is a set of regular participants in the game, $P = \{P_1, \ldots, P_m\}$. The participants produce verbal turns $T = \{t_1, \ldots, t_k\}$. In our setting, we assume that there is one special token that is used to flag a verbal turn $T$ as a proposal for an answer $A$ and another token to flag a turn as a mutual agreement on it; this type of game could hence also be called an *Agreement Game*.

Each solved task—that is, each annotation—can be represented as a tuple $\langle Q, I, A, T \rangle$. Our hypothesis is that the provided answers $A$, relative to the given information $I$ and the respective question $Q$, are of higher quality in settings where $T \setminus A$ is non-empty compared to those where it is; (that is, where there has been interaction between the annotators) and moreover, that the turns $T \setminus A$ in the interactive case provide insights into the reasoning steps that are taken to perform the mapping from $I$ to $A$, given $Q$—from which ultimately strategies for providing explanations could be learned.

## 3 Experiment

To test the hypotheses set out above, we created a number of tasks (pairs of questions $Q$ and information $I$), which we put to individual annotators and also to pairs of annotators in a dialogical setting.

### 3.1 Example Tasks

**Birds** Here, we show images of birds of two different kinds, as in Figure 1. The task for the annotators is to produce a characteristic description of one of the two kinds; i.e., a description that is true for all and only the images in the specified row. Following the question $Q$ = "what separates the birds in 1 from those in 2" given the images $I$ in Fig. (1) $A$ = "large wingspan, grey plumage with black head" would be a valid answer. The images are taken from the *Caltech-UCSD Birds 200-2011* dataset (Wah et al., 2011). Details on the setting can be found in the appendix.

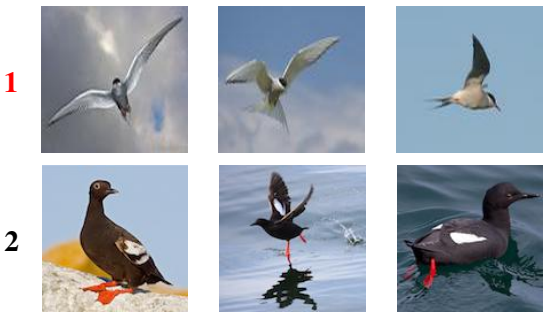


Figure 1: An Example Birds Task

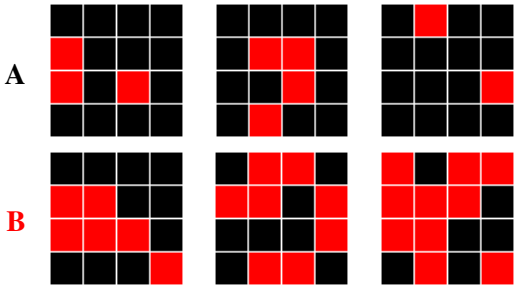


Figure 2: An Example Blocks Task

**Blocks** This task consists in providing a characteristic rule for one of two artificial, programmatically-created categories in the form of blocks with patterns. A valid answer for the example in Figure 2 could be "B has six or more red blocks." Note that in this kind of rule induction task from few examples, there will always be a large number of rules that correctly describe the pattern, even if they are different from the one that was actually used to generate the examples.

1 Daniel grabbed the milk there.
2 Sandra journeyed to the garden.
3 Sandra picked up the football there.
4 Sandra put down the football.
Where is the football?

Figure 3: An Example Texts Task

**Texts** To provide some variety, we also tested a text comprehension task, where a question about a text has to be answered; see Figure 3.

### 3.2 The Technical Setup & Collection

We realised the dialogical game interface as a web application, built on top of the chat server

*slurk* (Schlangen et al., 2018). The Mechanical Turk platform was used to recruit workers. After having read the instructions for the task, workers that accepted the task were transferred to a "waiting area" in the chat tool; as soon as a second worker entered this area, both the players were then moved to their task room (see also figure 4 in the appendix). The participants were paid an amount of $0.14 per minute (for a maximum of 4 minutes per game, although they could discuss longer). We also paid a bonus amount of $0.10 when the participants talked about things related to the task, tried to contribute equally during the discussion and also found the correct answer.

Additionally, we ran a monological version of the tasks with individual annotators, where we just presented the annotation task and collected the answer.

We collected 40 dialogues per setting, for a total of 120. Each dialogue consists of the consequtive discussion of two questions. After removing failed dialogues (where one participant left in the middle of the game, or participants clearly failed to follow the instructions), we were left with 93 dialogues: *28* for *birds*, *33* for *synthetic*, *32* for *text*. For monological annotation, we collected 40 annotations per setting, for a total of 120 annotations.

# 4 Results

## 4.1 Descriptive Overall Statistics

Table 1 shows some statistics about the collected data. In case of the dialogues, since the answers were marked by a prefix `/answer`, we could automatically identify them and look at the *discussion* (everything but the answer) and the *answer(s)* separately. "Speaker contribution ratio" is a measure of how balanced the dialogue was in terms of contributions by each participant. It is the ratio between the number of tokens produced by the more talkative participant and the number of tokens produced by the other participant; a perfectly balanced dialogue would rate 1 here. We also looked at the ratio of turns by each speaker.

As these numbers show, the participants in the dialogues produced more tokens overall, and, for Birds, also longer answers. The dialogues tended to be dominated by one speaker. When taking out the outliers (ratio above 3.4), which were cases where one participant had to explain the task to an inattentive other player, the imbalance is lower, but still pronounced, whereas it is not as strong in terms of turns. The numbers for Blocks and Texts are impacted by the high proportion of dialogue without any discussion (just `/answer` followed by `/agree`), as shown in row "no discussion dlgs". Looking deeper into the dialogues, we found that in about 65% of the cases, the more dominant speaker was also the one who proposed the final answer.

| Averages | Birds | Blocks | Texts |
|---|---|---|---|
| length (mins) | 5.25 | 5.87 | 5.63 |
| # turns | 4.30 | 3.39 | 2.86 |
| # turns w/o $A$s | 2.96 | 1.83 | 1.39 |
| # tokens | 39.61 | 28.09 | 18.33 |
| # tokens, final $A$ | 14.43 | 11.41 | 7.48 |
| speaker contr. ratio | 3.46 | 5.53 | 6.13 |
| ... w/o outliers | 2.85 | 4.15 | 4.30 |
| speaker turn ratio | 1.13 | 0.86 | 0.97 |
| no discussion dlgs | 35.7% | 56.1% | 57.8% |
| # tokens (monological) | 11.45 | 12.52 | 9.45 |

Table 1: Statistical Overview of Data

## 4.2 The Answers

While we can automatically identify the proposed answers by the players, we cannot automatically evaluate them. For Birds and Blocks, a wide variety of answers could be considered correct; for Texts, there is a single correct answer, but different ways of phrasing it. Hence, we manually classified the answers as *correct* and *incorrect*.

Incorrect answers often betray a misunderstanding of the task, as with "The birds in Section 2 look like the same type of bird, or breed. The birds in Section 1 all look like different types of birds, or breeds" for Birds, or "Mary is not in the bathroom because the statement is in past tense" for Texts.

Table 2 shows the accuracy of the final answers across tasks and settings (dialogue and monologue). The accuracy is measured by comparing the 40 answers $A$ to the corresponding 40 identical questions $Q$ for each task used for the dialogues and monologues. These results indicate that the tasks seem to be of different difficulty, with Birds eliciting the highest number of correct replies, and the constructed, quite challenging Blocks task the least, across settings. The numbers for the monological setting are consistently lower, lending support to our hypothesis that the dialogical setting leads to improved quality in the answer.

| Tasks | *Correct Answer*(%) | |
|---|---|---|
| | Dialogue | Monologue |
| Birds | 92.5 | 85.0 |
| Text | 90.0 | 85.0 |
| Blocks | 57.5 | 50.0 |

Table 2: Dialogue vs. Monologue: Correct Answers

### 4.3 The Discussions

To further analyze the discussions, we first categorized them as *active* or *passive*. In an active discussion parts of the final answer is "rehearsed" before the official reply is given or the final answer is assembled out of several turns. The following is an example of this category (for the task shown in Figure 1 above).

(1) A: Looks like the birds under 2 have red-orange feet.
B: The difference that I notice is that the birds in Section 1 are light feathered vs. the dark feathered birds of Section 2.
A: Ah, I like your answer better than mine.
B: */answer* The birds in section 1 do not have red-orange feet like the birds in section 2. Also, the feathers of the birds in Section 1 are light-colored vs. the dark-colored feathers of the birds in Section 2.
A: */agree*

We consider all other dialogues as passive. This includes cases where a proposal was immediately made and accepted, as well as where one partner didn't engage with the proposals. 28.6% of the Birds dialogues were passive, compared with 61.5% for Text and 65.5% for Blocks. This again shows an influence of the task; presumably, Text was considered too easy to warrant discussion, while Blocks may have been seen as too hard, with participants giving up (as also reflected in the accuracy on that task).

To unveil the reasoning steps of the collaborative thinking process we identified and quantified typical patterns in the active discussions. (2-a) shows an example of *Proposal Extension*, where a proposal made by A is implicitly accepted and extended; and of *Counter Proposal*, where a proposal is implicitly rejected and replaced with a counter proposal. (There were also explicit acceptances and rejections, w/o proposals.)

| Tasks | *Patterns in active dialogues*(%) | | | |
|---|---|---|---|---|
| | Proposal-Extension | Counter-Proposal | Explicit Acceptance | Explicit Rejection |
| Birds | 52 | 68 | 60 | 8 |
| Text | 80 | 40 | 60 | 0 |
| Blocks | 30 | 80 | 70 | 30 |

Table 3: Proportions of active dialogues in each task with different patterns.

(2) a. *Proposal Extension*
A: One obvious difference that I see from the birds in section 1 is that the birds have longer beaks. [Proposal]
B: another thing I noticed is it looks like 2's have softer feather colors [Proposal-Extension]
b. *Counter Proposal*
A: */answer* section 1 birds all look gray feathered [Proposal]
B: They all have yellow bodies and dark heads [Counter Proposal]

Table 3 shows the proportions of active dialogues in which these patterns were observed, by task type. Explicit rejections happened rarely but were never observed in the texts task (too simple task). For Birds, there seems to be a balance between proposal-extension, counter-proposal and acceptance (balanced discussion). There were fewer counter-proposals in texts task, for it being simpler. It also looks like there were more disagreements in Blocks due to the complexity of the task.

## 5 Conclusions

We have presented a setting for collecting annotations from pairs of interacting annotators. Our analysis indicates that this setting of an "agreement game", where explicit proposals have to be explicitly agreed on, fosters dialogs between the annotators. These dialogues yield to more correct responses and provide explication of the reasoning steps behind an annotation decision. Hence, both of our hypotheses, that the collaboration yields to more accurate responses and can reveal, at least in parts, the underlying reasoning steps, are supported. In line with our third and final hypothesis, the presence of these reasoning steps shows that the setting moves explanation as a linguistic act in the focus. It does however appear to be important to tune the level of difficulty of the task: if it is too simple, discussions do not emerge; if it is too hard,

the incentives for crowd workers have to be properly set so as to engage them. Our set-up also illustrates that natural categories could bring in more balanced discussions as well as better quality answers. Overall, it could provide useful data for developing a system which provides justifications.

## Acknowledgements

We gratefully acknowledge help with setting up the experiment from our Bielefeld student research assistant Ayten Tüfekci.

## References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An approach to evaluating interpretability of machine learning. *CoRR*, abs/1806.00069.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*.

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, New York, NY, USA.

Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge, UK.

David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarrieß. 2018. slurk - a lightweight interaction server for dialogue experiments and data collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / semdial)*, Aix-en-Provence, France.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Hao Yuan, Yongjun Chen, Xia Hu, and Shuiwang Ji. 2019. Interpreting deep models for text analysis via optimization and regularization methods. *AAAI Conference on Artificial Intelligence*.

Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. 2019. Interpreting cnns via decision trees. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

## A The Annotation Game Interface

Once two workers were presented they entered the task room as shown in Figure 4.

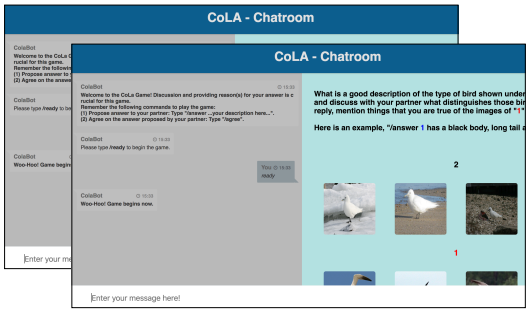


Figure 4: As soon as there are two participants in the waiting room, they are moved to the game room. Both participants see the same content on their screen. The content includes question $Q$, information $I$ and instructions by the game bot who is also present in the game room.

This setup technically realises the setting described formally in Section 2 above, where annotators ("participants") can work together to jointly formulate an answer $A$ to the question $Q$ they are given.