# Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation


Vivian Lai
vivian.lai@colorado.edu
University of Colorado Boulder
Boulder, CO, USA

Samuel Carton
samuel.carton@colorado.edu
University of Colorado Boulder
Boulder, CO, USA

Rajat Bhatnagar
rajat.bhatnagar@colorado.edu
Amazon
Seattle, WA, USA

Q. Vera Liao*
veraliao@microsoft.com
Microsoft Research
Montreal, Canada

Yunfeng Zhang†
yunfengz@twitter.com
Twitter Inc.
New York, NY, USA

Chenhao Tan
chenhao@uchicago.edu
University of Chicago
Chicago, IL, USA



## ABSTRACT

Despite impressive performance in many benchmark datasets, AI models can still make mistakes, especially among out-of-distribution examples. It remains an open question how such imperfect models can be used effectively in collaboration with humans. Prior work has focused on AI assistance that helps people make individual high-stakes decisions, which is not scalable for a large amount of relatively low-stakes decisions, e.g., moderating social media comments. Instead, we propose conditional delegation as an alternative paradigm for human-AI collaboration where humans create rules to indicate trustworthy regions of a model. Using content moderation as a testbed, we develop novel interfaces to assist humans in creating conditional delegation rules and conduct a randomized experiment with two datasets to simulate in-distribution and out-of-distribution scenarios. Our study demonstrates the promise of conditional delegation in improving model performance and provides insights into design for this novel paradigm, including the effect of AI explanations.



## CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Law, social and behavioral sciences**.

**ACM Reference Format:**
Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation . In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3491102.3501999


*Part of this work was completed while the fourth author was working at IBM Research.
†Part of this work was completed while the fifth author was working at IBM Research.




*CHI '22, April 29-May 5, 2022, New Orleans, LA, USA*



ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3501999


## 1 INTRODUCTION

As AI performance grows rapidly and even surpasses humans in benchmark datasets [11, 40, 49, 62, 76], AI models hold great promise for improving human decision making in a wide variety of domains. However, full automation may not be desirable for ethical, legal, and safety reasons, especially in high-stakes domains [14, 37, 51, 57]. In particular, one well-known problem with the current AI models is *distribution shift*. Namely, AI performance can significantly drop for out-of-distribution examples that are different from the training data (in-distribution examples) [9, 24, 46, 61].

Human-AI collaboration is thus critical for effective integration of AI models into human decision making processes [4–6, 8, 15, 63, 66, 82]. Many studies have investigated the role of AI in assisting humans in making individual decisions [9, 16, 36, 37, 50, 51, 55, 58, 69, 83, 84, 89], e.g., predicting whether a person will recidivate in the near future. Such decisions are non-trivial even for human experts (e.g., judges) and AI models can potentially offer insights through their predictions and explanations. This approach is well suited for high-stakes domains, where humans are expected to make the final decision on every case (e.g., judges in bailing decisions). However, human-AI collaboration on every single decision is not scalable and is thus less appropriate for tasks involving a large amount of relatively low-stakes decisions. One such example is content moderation, where moderator decisions on individual comments for further actions (e.g., hiding the content or prompting further review, depending on the community policy) are of limited consequence; instead the key challenge lies in dealing with the massive scale of comments. Such tasks can benefit from a greater level of automation [18, 33, 35].

In this work, we propose an alternative paradigm of human-AI collaboration — conditional delegation. Fig. 1(A) illustrates a general form of conditional delegation. Human and AI work together to identify trustworthy regions of AI before deployment, i.e., model decisions are reliable or trustworthy for examples within these regions. Once deployed, the AI model only affects decisions for instances in the trustworthy regions. For the rest, another set of actions can be taken such as manual review or employing a different model since the given AI's decisions on them cannot be trusted. This approach employs a greater level of automation than human-AI collaboration on every single decision and provides human with active control on when to use an AI model and in what ways.

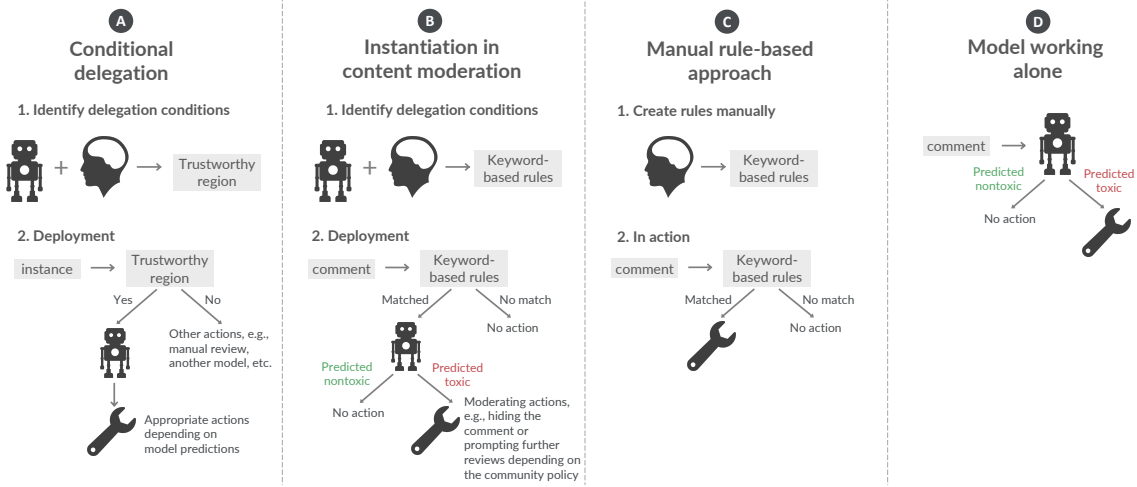



**Figure 1: Illustration of conditional delegation. Part A shows a general form of conditional delegation. Humans and AI work together to identify trustworthy regions of AI. Then once deployed, the AI model only affects the instances that belong to the trustworthy regions. Part B instantiates conditional delegation in the context of content moderation for this work. The right columns shows the contrast of the current manual rule-based approach for content moderation (Part C) and the model working alone (Part D).**

We use content moderation as a testbed. Fig. 1(B) shows one possible instantiation in this context. Trustworthy regions can be operationalized with a collection of keyword-based rules created by human-AI collaboration before deployment. For example, after inspecting AI predictions on comments with the word "retard", the human may decide that AI works well on them and set "retard" as a conditional delegation rule. Once deployed, comments that fall within these trustworthy regions, i.e., *containing any keywords* specified by human, if *predicted toxic*, can be reliably reported for final actions, such as being hidden or sent for further review, depending on the community policy.

Notably, the task for humans to create *conditional delegation rules* share some similarity with what many social media moderators are already doing by writing manual automation rules to deal with the massive amount of comments (Fig. 1(C)). For example, moderators on Reddit use a tool called AutoModerator, with which they manually customize a rule-based system to automatically identify comments for deleting or reporting for further review [18, 44]. This approach, however, misses out the benefit of AI especially since rigid rules often do not work on informal languages such as social media posts (e.g., containing swear words without being toxic). Without significantly altering content moderators' workflow, conditional delegation offers a promising approach to utilize AI, even if the model is not optimized for the community-specific content and should not be blindly trusted to work alone for every comment (Fig. 1(D)).

In this instantiation, a key difference from individual human-AI decision making lies in the success criteria: while the quality of individual decisions (e.g., accuracy) is often the target in individual decision making, *precision* and *coverage* are critical for conditional delegation because moderation actions will only happen on comments that are predicted toxic.[1] Precision ensures that AI behavior is indeed trustworthy in the delegation mode and avoids unnecessary actions, whether it is mistaken deletion or extra work for further review. Coverage warrants that the AI model can identify as many toxic comments as possible to alleviate the scalability issues. In the context of content moderation, recall (identifying all toxic comments) is often less of a priority given the limited time for content moderators, who are often volunteers, to deal with a massive amount of incoming comments. This is reflected in the current workflow using the manual rule-based approach (Fig. 1(C)), where comments falling outside the rules are ignored without taking an action. We assume the same workflow in our study and only focus on the precision and coverage related metrics for comments within the scope of keywords rules.

In this study, our *primary* interest is to investigate whether humans can effectively identify trustworthy regions for conditional delegation to improve the model precision with a good coverage, compared to the current manual rule-based approach (Fig. 1(C)) and the model working alone (Fig. 1(D)). Furthermore, we explore the effectiveness in two different AI scenarios: using an AI trained on the community specific data (*in-distribution*), and one trained on different data (*out-of-distribution*). The out-of-distribution model would perform much worse, but conditional delegation offers a potential means to improve through human-AI collaboration.

Our second set of contribution is to inform design of interfaces that support people to create high-quality conditional delegation rules. When given an AI model, content moderators often do not have labeled comments to quantify model performance. It would be helpful for them to observe model behaviors on their own data of interest to identify good delegation rules (i.e., trustworthy regions). To facilitate the creation of keyword-based rules, we develop an interface that allows participants who act as moderators to perform keywords search and observe model behavior on the search results. We provide and study the effects of several delegation support features, including predicted labels, local explanations that show

[1]Depending on the workflow, avoiding false negatives could be important in other instantiations.

the rationales behind predictions, and global explanations that provide an overview of the model.

To summarize, we ask the following research questions:

RQ1. Can users create keyword-based rules for conditional delegation that improves model precision, so that these rules correspond to trustworthy regions?

RQ2. How do the performance of conditional delegation and user experiences (such as engagement and subjective perceptions) vary between in-distribution and out-of-distribution AI?

RQ3. What are the effects of delegation support features on performance and user experiences, including showing prediction labels, local explanations, and global explanations?

Through a randomized experiment with 240 mechanical turkers, we show that even crowdworkers are able to create high-quality rules that lead to higher precision with conditional delegation than the model working alone. Especially when applied to an in-distribution AI, which already outperforms the manual rule-based approach for content moderation, conditional delegation further enhances the performance, leading to "complementary performance" (i.e., human+AI > AI and human+AI > human) [8]. For out-of-distribution AI used in this study, conditional delegation improves the model performance but does not suffice in compensating for the performance disadvantage of AI to outperform the manual rule-based approach. We also found that model explanations can improve efficiency in identifying delegation conditions and, with weak evidence, improve user experiences.

Overall, our work provides a new perspective to the emerging area of human-AI collaboration. Our core contribution is to demonstrate that conditional delegation is a promising alternative paradigm that allows users to control when to trust or distrust AI. We also contribute a set of interface features to assist people in creating conditional delegation rules and an empirical understanding of their effects. The diverging performance of in-distribution and out-of-distribution highlights the importance of considering the effect of distribution shift when conducting empirical studies of human-AI collaboration to inform the generalizablity of results, echoing recent findings in other studies [23, 56].

## 2 RELATED WORK

### 2.1 Human-AI Collaboration

Terms like "human-AI collaboration" [4, 5, 15, 82], "human-AI partnership" [63], "human-AI teaming" [6, 8, 66] have emerged in various literature studying the use of AI systems. They reflect a shift of perspective away from complete automation by AI. Fostering effective human-AI collaboration is not only critical for safety reasons, especially in high-stakes domains [14], but also necessary to harnessing the complementarity of human and AI intelligence to achieve optimal outcome [7, 86], reduce computational complexity [42], and enable novel technologies that are beyond the current capabilities of AI [25, 82].

Many forms of human-AI collaboration have been explored. The term "human-in-the-loop" is used broadly, but often refers to interactive training paradigm where the AI receives input from the human to improve its performance. For example, the field of interactive Machine Learning [3, 29, 30, 42], at the intersection of ML and HCI, develops systems that allow end users to guide model behavior. This kind of paradigm allows humans to directly impact the working of AI, and requires using algorithms that can incorporate human input to update the model, which can be technically challenging or infeasible in practice.

Another rich area to study human-AI collaboration is AI-assisted [13, 83, 89] or "machine/algorithm-in-the-loop" decision-making [37, 51]. In this paradigm, AI performs an assistive role by providing a prediction or recommendation, while the human decision maker makes the final call and may choose to accept or reject the AI recommendation. Several studies explored the questions of whether and how to achieve *complementary performance*, i.e., the collaborative decision outcome outperforming human or AI alone [8, 51, 89]. The empirical results, however, are mixed at best, because there was either insufficient complementarity in human and AI's domain knowledge or a lack of ability for people to judge the reliability of AI recommendations. This approach tends to focus on high-stakes decisions and are not scalable in the number of decisions because humans are required to make each decision.

Another line of work explores intelligent systems and considers different tasks that AI can perform and the optimal level of automation versus human agency [51, 59, 81]. For example, building on a classic model of levels of automation [67], Mackeprang et al. [59] proposed a design framework that decomposes the design space of an intelligent system into sub-tasks then allocates human, AI or both to perform each sub-task.

The goal of our work is to have AI partially automate a large volume of decisions rather than assisting individual decisions. Extending existing models of human agency and automation [59, 67], we introduce *proactive* human agency, with which human can act and exercise control prior to model deployment, instead of reacting to model outputs. By conducting a controlled experiment, we explore whether this new human-AI collaboration paradigm can achieve complementary performance by outperforming AI and manual approaches. While some prior work also discussed delegation based on predicted outputs (e.g., predicted probability) [18, 47], our work focuses on identifying trustworthy regions in the input space. Furthermore, to the best of our knowledge, our work is the first study with controlled experiments to examine the effect of conditional delegation.

### 2.2 AI explanations for human-AI interaction

Mental model, defined as an understanding of how a system works, is a key concept in human-computer interaction [65]. Having an appropriate mental model allows people to accurately anticipate a system's behaviors and interact more effectively. People's mental model can be refined by explanations of how the system works. Therefore, explanation and transparency features have long been an interest of HCI research on various technologies [1, 41, 54, 71].

Recently, AI explanations have gained much attention [8, 13, 27, 32, 51, 53, 89]. The popularity of complex, inscrutable AI models such as deep neural networks make the difficulty of understanding a primary challenge for modern AI technologies. This challenge has given rise to a technical field of explainable AI (XAI), producing an abundance of techniques that aim to make AI more understandable by people. While the landscape of XAI technique is beyond the scope of this paper [2, 34, 38], an important distinction relevant to

our study is the contrast between *local explanations*, which focus on explaining the rationale for a particular prediction, versus *global explanations*, which aim to give a high-level understanding of how the AI works. We explore the effect of both types of explanation in our study and will discuss the details of the XAI techniques used for our toxicity prediction model in the next section.

HCI studies on XAI have found explanations to improve user understanding of AI systems [12, 22, 32], and somewhat mixed results on enhancing user trust [22], satisfaction [32] and willingness to adopt AI systems [80]. Moreover, explanations provide additional information that can be utilized to assist the task that people perform. For example, Lai and Tan proposed a spectrum between human agency and full automation for machine learning to assist human decision-making [51], and considered showing explanation as an additional form of machine assistance beyond solely providing prediction labels, and thus increase the level of automated assistance. In interactive machine learning, explanation has been studied as a primary means for people to directly inspect the model limitations, instead of just observing model behaviors, for people to provide feedback [32, 79] to improve the model.

For our conditional delegation task, we hypothesize that explanations of AI model predictions, i.e., keywords that the model bases its prediction on, can give hints to people about keyword rules they should consider, and potentially help them judge the effectiveness of a given rule.

## 2.3 Distribution Shift and Experimental Studies on Out-of-distribution Examples

Current AI models rely on identifying patterns in training datasets. In a real-world scenario, it is unlikely that models are used to classify data that is exactly the same as the training dataset. For instance, a moderation team would likely work with a model trained on an existing dataset, then applied to the data on their platform. The difference between the training dataset and the deployment data is called distribution shift, which often results in a performance drop [24, 46, 61]. For instance, McCoy et al. [61] find that state-of-the-art models in natural language inference adopt three fallible syntactic heuristics and perform around random chance when tested on examples where these heuristics fail.

Despite substantial interest in distribution shift in the AI community, the effect has been rarely examined in empirical studies of human-AI collaboration, with a few recent exceptions [23, 56]. Liu et al. [56] demonstrated that there exists a clear difference between in-distribution and out-of-distribution examples when human and AI collaborate to make individual decisions in recidivism prediction and profession prediction. They suggested that complementary performance is more plausible for out-of-distribution examples because of AI's performance drop. Chiang and Yin [23] examined human reliance on the model in human-AI decision making and found that surprisingly humans rely on AI more out-of-distribution, where the AI performance is worse.

The existence of distribution shift is a strong motivation for some form of conditional delegation so that humans can identify the trustworthy regions. In our setup, however, as we conditionally delegate decisions to AI, strong AI performance in-distribution is likely more critical for the human-AI collaborative performance. We thus hypothesize that it is more challenging to identify the trustworthy regions for out-of-distribution examples because the model behavior is likely more spurious.

## 2.4 Content Moderation

Content moderation has attracted substantial interest from the research community due to its growing importance in online communities [48]. There is a large body of research studying the effect of moderation on community behavior, including whether one should regulate at all [17, 19, 21, 45, 75, 78]. In contrast, our work is concerned with the practice of content moderation, i.e., how moderators can efficiently deal with a large number of comments. The scale of content is the most important argument for some form of automation in content moderation [33, 35]. Moreover, an active line of research has investigated the "emotional labor" of moderation work by the volunteer moderators [28, 60, 72], further highlighting the importance of avoiding burnout for moderators through automation.

One strategy is to use rule-based methods. For instance, Reddit moderators can configure an AutoModerator bot to set rules for reporting or deleting all comments that contain certain words.[2] The key advantage of this method is that it is entirely under the control of moderators. Through interviews with 16 moderators, Jhaver et al. [44] found that AutoModerator improves the efficiency of moderation. However, there exists a need for audit tools to monitor the performance of the keyword rules. They also highlight the fact that AutoModerator fundamentally changes the work of moderators and may introduce additional unnecessary work. Chandrasekharan et al. [18] also found that hard-coded rules are prone to mistakes.

An alternative strategy is to use AI models beyond rule-based approaches. Toxic comment detection or hatespeech detection has attracted a lot of interest from the AI community [64, 70, 85, 88]. Notably, the Perspective API is reportedly used by the New York Times, Disqus, and other platforms.[3] However, researchers increasingly recognize the pitfalls of full automation: 1) models are trained with historical data and can present issues such as gender bias and racial bias in AI models [68, 73], potentially exacerbating structural inequalities [10]; 2) there exist diverse rules and preferences of austerity and value in different communities [20, 31, 74, 77]. Anecdotally, we deployed a version of our model on a subreddit to report comments that are predicted as toxic, and the moderators asked us to shut it down due to high false positive rates (i.e., low precision). Inspired by the diversity of rules, Chandrasekharan et al. [18] proposed a new system that combines classifiers based on different communities and advocated that this tool be configured as part of moderation workflow.

Our effort represents a new direction in exploring the mixed initiative in content moderation. Conditional delegation combines traditional rule-based approaches and AI models by providing moderators with the ability to decide when to trust or distrust the AI model. Such rules can be created for any model of choice, so it is orthogonal to the research on improving the capability of AI. It can also be used to tailor different requirements of precision and tune the tradeoffs between false positives and false negatives.

[2] https://www.reddit.com/wiki/automoderator.
[3] https://www.perspectiveapi.com/case-studies/.

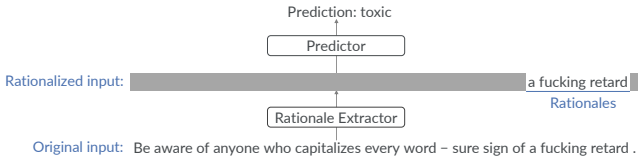



**Figure 2: Illustration of the model with an example. The rationale extractor first identifies "rationales" in the input, and then the predictor makes the prediction based on the rationales. This model can achieve competitive accuracy while having built-in interpretability because the prediction is made exclusively based on rationales.**

## 3 AI MODEL

A critical component of our study is the model used to assist people in content moderation. In this section, we present details of how we obtain the model used in our study and provide an overview of its properties.

### 3.1 Model Development

Current AI models are driven by the data used to train the model. We choose two datasets to simulate the in-distribution and out-of-distribution scenarios. We then develop an interpretable model that is trained on the in-distribution data and achieves reasonable performance on the out-of-distribution data.

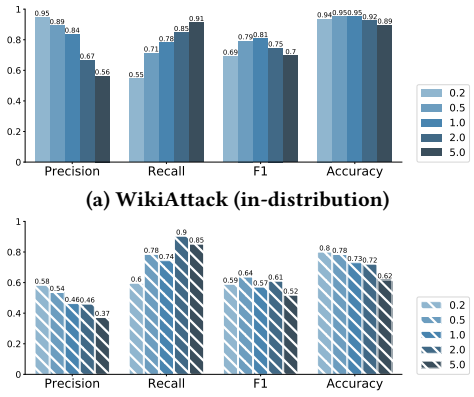
**Data.** In this work, we use a dataset of Wikipedia comments [88] (henceforth *WikiAttack*), made public by Wikipedia and Google Jigsaw. Notably, Jigsaw powers the Perspective API[4], a popular free service for toxic comment detection. Therefore, using a model derived from this dataset allows ecological validity to our study as the dataset is used by real-world social media platform and community moderators. We use the original train/test split of Wulczyn et al. [88], resulting in 70k comments in the training set and 23K comments in the test set. We use the test set to evaluate the ability of participants to create keyword-based rules for conditional delegation for *WikiAttack*.

To simulate the out-of-distribution scenario, we use another dataset of hate speech on Reddit [70], consisting of 22K comments, on which we apply the same model mentioned above. As a result, the datasets that participants explore to create rules are of comparable size between Wikipedia (in-distribution) and Reddit (out-of-distribution). Throughout the rest of the paper, we will use *in-distribution* and *WikiAttack*, *out-of-distribution* and *Reddit* interchangeably.

**Model.** We use a rationale-style neural architecture [52] as the classifier underpinning our tool, producing both explanations and predictions. Fig. 2 illustrates our model architecture. It uses one text encoder to identify rationales (i.e., a subset of tokens) from the input, and another text encoder to make predictions based on the rationales.[5] Trained in tandem with a sparsity objective on the rationales, this model attempts to obscure as much of the input as possible while still leaving enough to make an accurate classification. In short, this model achieves competitive accuracy while having the ability to provide explanations directly by showing the rationales on which the prediction is based on.

[4]https://www.perspectiveapi.com/

[5]Technically, the predictor uses the masked input.




**(a) WikiAttack (in-distribution)**

**(b) Reddit (out-of-distribution)**

**Figure 3: An overview of model performance with different hyper-parameters. The hyperparameter shows the relative weight of recall vs. precision. We choose the model with 0.5 because it achieves competitive performance both in-distribution and out-of-distribution. Throughout this paper, we use solid bars for WikiAttack (in-distribution) and bars with white slashes for Reddit (out-of-distribution).**

For the generator and predictor, we use independent, pretrained BERT [26] instances distributed by HuggingFace [87]. We use Pytorch Lightning[6] for fine-tuning. We use Gumbel Softmax [43] to enforce a binary constraint on the predicted rationale, such that a token is either fully included or fully excluded from the input. As an implementation detail, we find it highly useful to pre-fine-tune the predictor layer on the full (un-masked) input before further training it in tandem with the generator.

Because our task emphasizes precision over accuracy, we experiment with different parameters to trade-off precision and recall. Fig. 3 shows model performance both in-distribution and out-of-distribution with different parameters. We observe a clear performance drop out-of-distribution (e.g., F1 drops from about ~0.8 to ~0.6), which validated our choice of Reddit as an out-of-distribution scenario. In our experiments, we choose the model with recall weight 0.5 (the second bar). Note that participants did not have access to this performance data because our goal is to simulate the scenario where moderators work with a model developed on an existing dataset. It is up to the moderators to figure out how well the model performs and when to trust or distrust the model.

This model can achieve BERT-like accuracy while being able to precisely and parsimoniously identify the rationale responsible for its prediction. Table 1 presents example rationales for comments that are predicted toxic, both correctly and incorrectly. We find qualitatively that the model produces sensible rationales in this application. While it identifies some surprising tokens as toxic such as "you", it does succeed in learning that the primary evidence of non-toxicity is a lack of toxic tokens: it retains only 2% of tokens on average for predicted-nontoxic comments, versus 15% for predicted-toxic comments. Note that this explanation method has attracted some criticism for producing rationales that don't necessarily align with human reasoning [90], but it has the advantage of producing rationales that are, by construction, sufficient (in the

[6]https://www.pytorchlightning.ai/.

| | |
|---|---|
| | WikiAttack |
| True positive | == Hello == Your such a fucking freak mate. Get a life. And get a facelift you ugly cunt. |
| True positive | == FUCKYOU == wow,just wow james delenoy. u are a class A faggot. i could pwn u in a coma. i dont give a fuck if u block this account cuz we are a freight train from hell and we will make more and come at u with the ferocity of a mama bear. SUCK OUR GRUNDLES, the entire waltmanist community |
| False positive | ::Haha, I fucking pissed myself reading this ') Thanks, i needed that after 3 hours trecking around in warrington with the worst cold ever. I mean it, thanks )) |
| | Reddit |
| True positive | Crawl back to you r/politics safe space. I'm sick of fucking faggot liberals coming into communities where they aren't welcome and trying to ruin them from the inside out. You're on KiA, so you already know how that works. Fuck you |
| True positive | Ah yes, coming from the idiot that calls people a retard when he disagrees. Get the fuck out of your internet bubble and look around you irl for a moment. |
| False positive | Someone in chat baited me into saying "paki". I got banned. I'm fucking Pakistani and that word has never been "offensive" to anyone. I've used it for 14 years myself. I had 1k hours in that game but it's staying uninstalled |

**Table 1: Comments that are predicted toxic from WikiAttack and Reddit with their identified rationales by our model.**

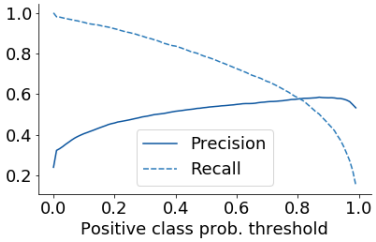
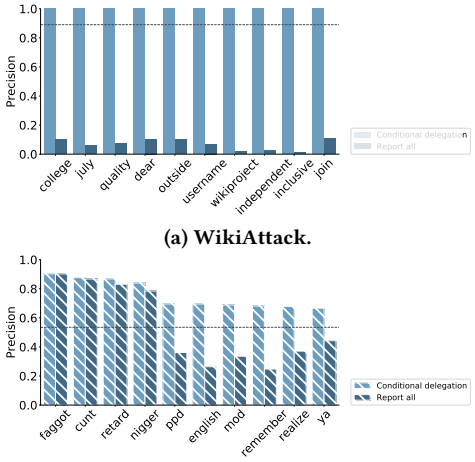



**Figure 4: Precision-recall plot of model on Reddit dataset (out-of-distribution). Even at high positive class probability thresholds, precision remains low.**




**(a) WikiAttack.**

**(b) Reddit.**

**Figure 5: Words with top precision on WikiAttack (in-distribution) and Reddit (out-of-distribution). "Conditional delegation" shows precision among comments with the word based on model predictions, while "Report all" shows this measure if we consider a comment toxic as long as it contains the word (manual rule-based approach). Dashed lines show model precision on all comments (i.e., the precision of the model working alone). In both cases, all the top 10 words lead to greater precision than the model working alone.**

logical sense) for the model's prediction. Generating high-quality explanations is an active area of research, and our paradigm of conditional delegation can be used for any model of choice.

The rationale produced by this model is a form of **local explanations**, identifying important words in each prediction. Our experiment also includes **global explanations**, which convey an overview of the model behavior across all inputs. We generate these global explanations by identifying the tokens that occur most frequently in the rationales of the model on the in-distribution and out-of-distribution data respectively. We display these top-15 most frequent rationale tokens (Table 2) as the "global explanations". We can immediately observe differences between Reddit and WikiAttack: "cunt" and "retard" are not common in rationales in WikiAttack but are among the top five on Reddit.

To provide context for our later findings, it is difficult to produce precise classification out-of-distribution. Fig. 4 shows that even with a high positive class probability threshold (0.93), the model only climbs to 0.58 precision. Thus, even a very conservative application of this model is still producing 2 false positives for every 3 true positives–impractical for use by real moderators, and something we would like to be able to improve on via conditional delegation.

## 3.2 Performance of Individual Words

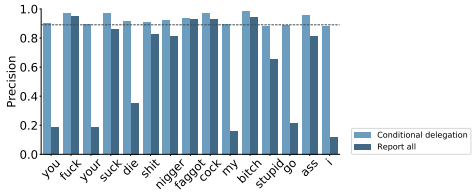
The goal of this work is to explore human-created keywords rule for conditional delegation to AI, such as "if a comment contains word X and is predicted toxic, the model will be trusted to report the comment for moderation action". In comparison, with a manual rule-based approach (e.g., the current AutoModerator system used by content moderators of Reddit), such a rule takes a form like "if a comment contains word X, that comment will be reported". Our hypothesis is that with the proper choice of rules, humans can produce a system which is more precise than either the manual rule-based approach or the model working alone.

We perform preliminary analysis to characterize the scope of the potential improvement and to contextualize our experimental results. A crucial question in motivating our approach is whether there exist trustworthy regions of the model, i.e., are there certain words that occur systematically in comments where the model achieves high precision.

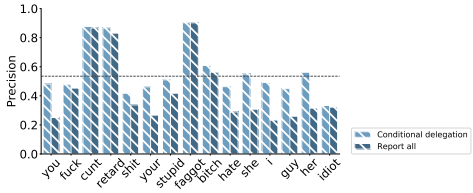
First, we compute the precision of conditional delegation for all words that show up in at least 100 comments. Fig. 5 shows the 10 words with the highest precision as conditional delegation rules

| | |
|---|---|
| WikiAttack | you, fuck, your, suck, die, shit, nigger, faggot, cock, my, bitch, stupid, go, ass, i |
| Reddit | you, fuck, cunt, retard, shit, your, stupid, faggot, bitch, hate, she, i, guy, her, idiot |

**Table 2: Words that are most frequently used in rationales.**




**(a) WikiAttack.**



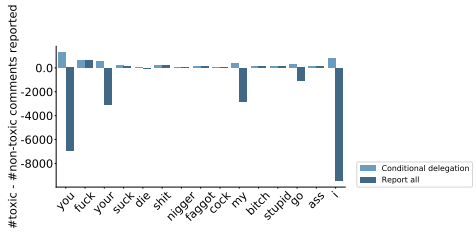

**(b) Reddit.**

**Figure 6: Precision for words that show up most frequently in rationales on WikiAttack (in-distribution) and Reddit (out-of-distribution) (ordered by frequency). "Conditional delegation" shows precision among comments with the word based on model predictions, while "Report all" shows this measure if we consider a comment toxic as long as it contains the word (the manual rule-based approach). Dashed lines show model precision on all comments (i.e., the precision of the model working alone).**

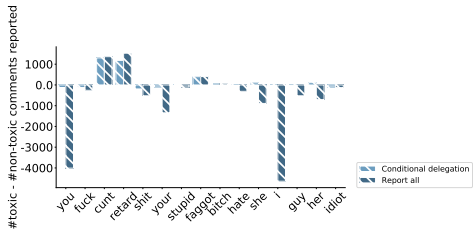
(i.e., based on model predictions for all comments containing them) on WikiAttack and Reddit respectively. Conditional delegation based on these words leads to greater precision than the model working alone (dashed lines), suggesting that users can improve the precision of the model by identifying these words for conditional delegation. In addition, we compare that with the precision of using the word as a "report all" rule as with manual rule-based approach, by considering all comments containing the word as toxic. We can see generally, for these words with top precision, trusting the model leads to higher precision than "report all", both in-distribution and out-of-distribution. However, the difference is much smaller for Reddit (out-of-distribution). In particular, "faggot", "cunt", "retard", and "nigger" achieve very high precision on this dataset even if one simply reports all comments that contain any of those words. These results indicate that conditional delegation can outperform both the manual rule-based approach and the model working alone if users are able to make good choices of keywords rules.

Next, we examine the precision of the words that are most frequent in rationales (Table 2, to be shown as global explanations). Fig. 6 shows that on WikiAttack, the majority of global explanations achieve greater precision than the model working alone. However, on Reddit, this is true only for six words ("cunt", "retard", "faggot", "bitch", "she", "her"), indicating the challenge of creating good rules for out-of-distribution data.

Fig. 7 further shows (the number of reported toxic comments - the number of reported non-toxic comments) if that word is chosen as a rule. This measure reflects both the coverage and precision of a




**(a) WikiAttack**




**(b) Reddit**

**Figure 7: Reward (number of reported toxic comments - number of reported non-toxic comments), a measure combining precision and coverage, for words that show up most frequently in rationales on WikiAttack (in-distribution) and Reddit (out-of-distribution) (ordered by frequency). "Conditional delegation" shows reward for comments with the word based on model predictions, while "Report all" shows this measure if we consider a comment toxic as long as it contains a word (the manual rule-based approach).**

keyword rule. We refer to this measure as *reward* because it is used as an incentive in our human subject experiments, to be introduced later. On WikiAttack, most global explanations lead to positive rewards. We also observe the clear advantage of using conditional delegation over "report all". This advantage becomes smaller on Reddit and disappears for "retard" and "cunt" because these two words have great precision and coverage by simply reporting all comments with them. In fact, rewards on Reddit are dominated by "retard" and "cunt" due to their high coverage. A user could achieve a quite high reward (and outperform the model) simply by reporting all comments with either of these two words.

In summary, there are specific words that delineate trustworthy regions of the AI model, and even certain words ("retard", "cunt") where simply reporting all comments containing these words would be more effective in terms of reward than delegating such comments to the model, particularly in the out-of-distribution setting. However, we are able to recognize such words with the benefit of a fully-labeled dataset (i.e., oracle access) — discovering them in a real-world setting could be very challenging. We explore this challenge in a rigorous human subject experiment in the next section.

## 4 EXPERIMENTAL DESIGN

Equipped with the model, one goal of this study is to design interfaces with different support features to enable people to come

up with effective keyword-based rules for conditional delegation. We then examine the effect of these support features through a human-subject experiment. In this section, we start by introducing different types of support features that we consider and then explain the study procedure.

## 4.1 Experimental Conditions and Interface Design

In order to enable people to create keyword-based rules for conditional delegation and observe model behaviors with them, the basic function of our tool is to search for a keyword and browse comments that contain it. This allows users to determine whether a keyword would serve as a good rule. For different experimental conditions, our design space mainly involves what information we provide when returning the search results.

**Experimental conditions.** As discussed in Section 3, in addition to predicting whether a comment is toxic or not, our model can provide local explanations (i.e., which words are used as rationales for the prediction) as well as global explanations (i.e., most frequent words that show up in the rationales). Therefore, we consider the following four conditions:

- **Predicted labels.** Predicted labels are shown along with the searched comments.
- **Predicted labels + local explanations.** In addition to predicted label for each comment, we highlight rationales, i.e., words in the comment the model uses to determine toxicity for comments that are predicted toxic. We refer to this condition as "*local explanations*".
- **Predicted labels + local explanations + global explanations.** Participants have access to all of the features in the previous condition and are also provided a list of words that the model typically uses in determining comment toxicity (Table 2). We refer to this condition as "global explanations".
- **Manual condition.** The final condition is designed to simulate the current state of AutoModerator, where moderators come up with "report all" rules. We create a consistent interface where participants have the ability to search comments and browse returned results to assess whether they are indeed toxic, instead of whether the model prediction is precise. Participants do not have access to any model-related information.

In the rest of this paper, we refer to these conditions as *experimental conditions* and WikiAttack vs. Reddit as *distribution types*.

**Interface design.** We start by introducing the interface for "Predicted label + local explanations + global explanations", which includes all possible components of the other conditions (see Fig. 8). The widgets in the interface are arranged in two columns, where instructions and comments are displayed on the left, while the search bar and the current set of rules are on the right. The instructions box (Fig. 8(1)) reminds participants of the task and provides more information about the interface to ensure that they can fully leverage the tool's features. Global explanations are shown below the instructions box (Fig. 8(3)). When the participant clicks on a rule that is represented by a button, it automatically searches comments with the respective keyword-based rule. In addition to searching particular words, we also allow users to load random comments, which can be used to explore the data (Fig. 8(4)). Upon a query or loading random comments, the comments are displayed as *cards* below the *load random comments* button. Depending on the condition, a comment could have a predicted label (Fig. 8(5)) and rationales could be highlighted (Fig. 8(6)).

On the right side, the first two widgets are the *search bar* (Fig. 8(7)) and *clear* button (Fig. 8(8)). The participant enters keyword-based rules and then comments with the respective rule are shown on the left, as described in the previous paragraph. Participants can filter comments by their predicted label (Fig. 8(9)). By default, both predicted toxic and nontoxic comments are shown. When the participant is satisfied with the rule, they may click on the *add rule* button (Fig. 8(10)) to add the rule to their list. All of the participant's rules are displayed in the component below the *add rule* button. We also display their total matched comments and predicted toxic matched comments (Fig. 8(11)). Finally, participants may click on the *finish making rules and go to survey* button to submit their rules and proceed to the exit survey (Fig. 8(12)).

Fig. 8 shows interfaces for the other three conditions. "Predicted labels + local explanations" condition (Fig. 9b) removes the global explanations (Fig. 8(3)) and is otherwise the same as "Predicted labels + local explanations + global explanations". "Predicted labels" condition (Fig. 9a) further removes the highlights of rationales when returning search results. "Manual" condition (Fig. 9c) only has "search" and "load random comments".

We conducted pilot studies with lab mates at the same university and mechanical turkers to gather feedback and ensured that interactions were intuitive and the interface was easy to use. With many widgets in the interface, it could potentially become overwhelming for the user. To overcome that problem, we included a short tutorial walk-through (Fig. 8(2)) that explains each widget in sufficient details. To minimize confusion, we also included a separate link in the instructions box (Fig. 8(1)) that enumerates how each widget should be used.

## 4.2 Study Design and Procedure

We employ a between-subject design with the four conditions described above. We aimed to obtain 0.95 power to detect an effect size of 0.1 in precision at the standard 0.01 alpha error probability using F-tests. As such, we recruited 30 participants for each condition, summing up to 120 participants each for in-distribution and out-of-distribution and 240 participants in total. The study is approved by the University IRB.

**Task flow.** Participants went through four phases during the study: (1) consent and introduction; (2) interface tutorial; 3) the actual task; (3) exit survey. During the introduction phase, participants were shown an accordion of instructions. Due to the nature of toxic comments, they are warned that the task will contain profanity and hate speech and are cautioned to only proceed if they are okay with reading toxic comments.

After the instruction, participants are asked three attention-check questions to allow us to filter out participants who did not pay attention to the instructions. They are given toxicity trial questions with four simple comments to label if they are toxic or nontoxic: (1) *I had a really bad day today.* (2) *I just made a fucking great steak!* (3) *Go fuck yourself.* (4) *I hope you die painfully tomorrow.* These questions allow us to filter participants who finish this task hastily

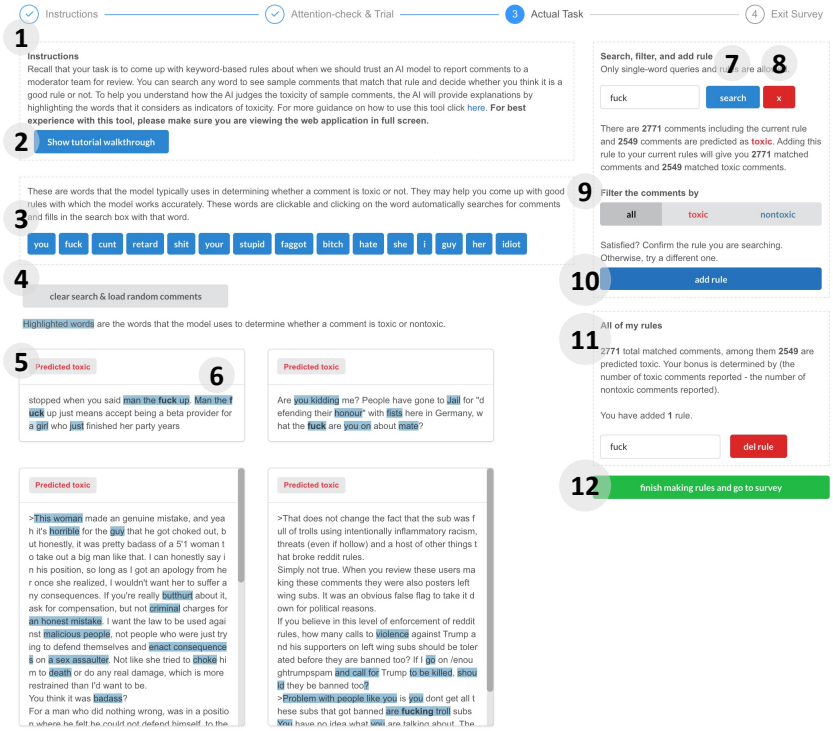
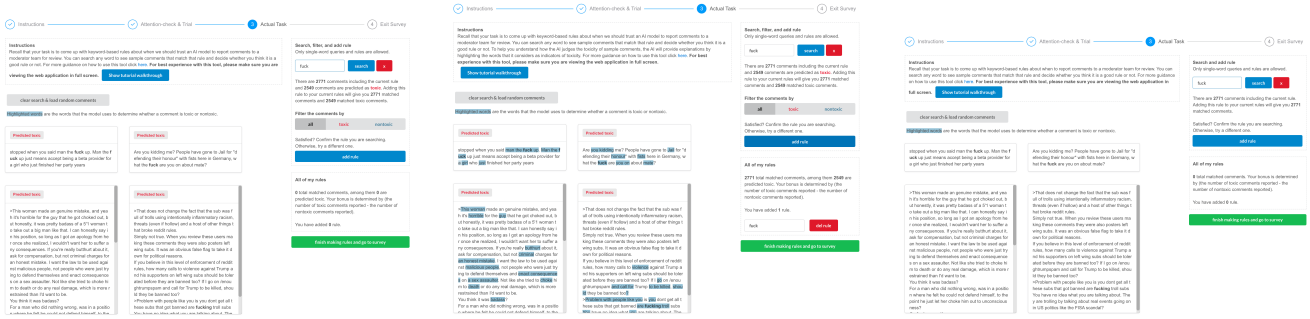



**Figure 8: Interface for "Predicted label + local explanations + global explanations". We use this interface to go through the design of all delegation support features.**



**(a) Predicted labels condition** **(b) Predicted labels + local explanations condition** **(c) Manual condition**

**Figure 9: Interfaces for the other three experimental conditions.**

without paying effort and prepare them for toxicity judgment. We remove participants whose accuracy is less than or equal to 50% on these questions. As a result, we filter 6 out of 240 in our analysis.

To familiarize participants with the interface, we include a tutorial walk-through when they first land on the page to instruct them on how to use each feature. We also include a link that featured more detailed instructions and a short demonstration video. Participants are required to submit at least ten rules, and can then exit the task whenever they are satisfied with the set of rules created.

In the exit survey, we collected basic demographic information, their knowledge and familiarity of AI and content moderation, and subjective measures, to be introduced in the later section.

**Reward.** To motivate quality work, in addition to a base payment, we design a bonus incentive as follows: participants will be awarded $0.10 for every 100 toxic comment their rules correctly reported, and penalize them $0.10 for every 100 nontoxic comment their rules mistakenly reported (lower bounded by $0 and upper bounded by $2). This bonus thus rewards both precision–how likely comments under the rule (for the manual condition) or conditional delegation with the rule are correctly classified as toxic, and coverage–the quantity of comments covered by the rule. To make the calculation easy to understand for participants, this reward makes a simplified assumption that the cost of wrongly reported non-toxic comments (false positive) equals the benefit of correctly reported toxic comments (true positive).

This reward mechanism is explained to participants, and we include one question in attention check to ensure they understood it. We also explicitly suggest that, to optimize for the reward, their

goal should be to come up with keywords that meet the following criteria: (1) that occur in a lot of comments; (2) with which the model makes accurate predictions on the comments, and (3) that are a diverse set so they may cover different kinds of toxic comments.

**Participant information.** We recruited participants from Amazon Mechanical Turk. We note that while this recruiting choice may limit the generalizability of our results, social media content moderation is often performed by part-time volunteers whose expertise varies. Furthermore, we believe turkers are a sufficiently good sample for us to compare whether conditional delegation improves the content moderation outcomes over the baseline condition, and expert users are likely to further enhance the improvement pattern, if any. We encourage future work to further test the paradigm in realistic social media contexts.

To ensure high quality responses, all participants satisfy the following criteria: (1) performed at least 1000 HITs; (2) approval of 99% performed HITs in previous requesters; (3) reside in the United States; 4) has the adult content qualification since our task shows toxic comments. The experiment follows a between-subject design therefore we do not allow any repeated participants.

There were 115 male, 116 female, 2 non binary, and 1 preferred not to answer. 52 participants are aged 18-29, 114 aged 30-39, 34 aged 40-49, 26 50-59, 7 aged over 59, and 1 I prefer not to answer. Participants rated their knowledge on artificial intelligence (25 had no knowledge, 156 had little knowledge, 49 had some knowledge, 4 had a lot of knowledge), and social media content moderation (36 had no knowledge, 113 had little knowledge, 66 had some knowledge, 19 had a lot of knowledge) on five-point Likert scales. Participants were paid an average wage of $11.80 per hour.

Overall, most turkers are satisfied with our task design and interface. Here are two quotes from their feedback: "*This was super intriguing. I had never participated in an activity like this before. It was hard coming up with bad words since they are not part of my vocabulary. It was interesting to see which words usually coincided with toxic subjects. Overall, very interesting project.*" and "*It was interesting. I see now how difficult moderation can be for some sites.*"

## 4.3 Evaluation Measures

We consider three types of evaluation measures to cover efficacy, efficiency, and subjective perception.

**Efficacy.** As discussed in Section 1, our main goal is to examine whether humans can improve the precision of the model with a good coverage via conditional delegation. We consider two precision-based measures: *average precision* and *union precision*. For the first three experimental conditions with delegation support features, average precision is formally defined as

$$\frac{1}{|R|}\sum_{r\in R}\frac{|\{x \text{ is toxic } \&\ x \text{ contains } r\ \&\ x \text{ is predicted toxic}\}|}{|\{x \text{ contains } r\ \&\ x \text{ is predicted toxic}\}|},$$

where $R$ is the set of rules that participants choose and $x$ refers to a comment, whereas union precision is formally defined as

$$\frac{|\{x \text{ is toxic } \&\ x \text{ contains any } r \in R\ \&\ x \text{ is predicted toxic}\}|}{|\{x \text{ contains any } r \in R\ \&\ x \text{ is predicted toxic}\}|}.$$

As the manual condition does not have a model, these two definitions become $\frac{1}{|R|}\sum_{r\in R}\frac{|\{x \text{ is toxic } \&\ x \text{ contains } r\}|}{|\{x \text{ contains } r\}|}$ and $\frac{|\{x \text{ is toxic } \&\ x \text{ contains any } r\in R\}|}{|\{x \text{ contains any } r\in R\}|}$.

The difference in the denominators highlights the role of conditional delegation, which only affect the comments that the model predicts as toxic. It follows that the performance with conditional delegation is also determined by the model's base performance, i.e., how well the model can identify toxic comments. Intuitively, average precision reflects the average quality of every single rule a person provides, while union precision measures the performance when using all rules from the person as a set, and can be skewed by the performance of higher-coverage rule in the set. Thus, one's ability to come up with both high-precision and high-coverage rules can lead to better union precision.

Finally, we consider the *reward* participants received, as introduced in Section 4.2, which measures the quantity difference between reported toxic comments (true positive) and reported non-toxic comments (false positive). This measure reflects both precision and coverage. This metric is highly volatile because a small number of keywords can achieve much higher rewards than others, especially out-of-distribution (e.g., "retard" and "cunt" on Reddit as shown in Section 3). We believe that precision is the more reliable measure of efficacy given that our participants tended to only choose about 10 rules.

**Engagement and efficiency.** We consider number of logged actions a participant took during the experiment task and number of rules they added as measurements for engagement. 13 types of unique actions were logged, including searching a rule, filtering comments by predicted labels (toxic and nontoxic), load random comments, get page comments, etc. We consider the number of actions more indicative of engagement, since participants can search for a rule without adding it. For efficiency, we consider total elapsed time and rules per minute. Elapsed time starts from the moment participants enter the interactive interface until they click on "finish making rules and go to survey", in minutes. Rules per minute is the number of rules added divided by elapsed time. Since rules are the final product of the task, rules per minute is more indicative of efficiency.

**Subjective measures.** Finally, we consider the following three categories of subjective perception, all gathered by the exit survey, using a five-point Likert scale (Strongly Disagree to Strongly Agree) for all scale items.

- Subjective workload. We adopt three applicable items from NASA-TLX [39]:
  - **Mental demand.** I felt that the task was mentally demanding.
  - **Feelings of success**. I felt successful accomplishing what I was asked to do.
  - **Negative emotions.** I was stressed, insecure, discouraged, irritated, and annoyed during the task.
- Confidence. There are multiple loci of confidence in this task: in the model, in one's own ability to create conditional delegation rules, and in the human-AI collaborative outcome. So we consider the following three measure (they were not asked in the manual condition since they do not apply):
  - **Confidence in model.** I trust the model to be able to correctly identify most toxic comments.
  - **Confidence in created rules** I am confident that my rules significantly improve the model's accuracy in detecting toxic comments.

  - **Confidence in deployment.** I am confident that my moderator team would feel comfortable relying on the AI model combined with the rules I provided.
- Understanding. We are interested in whether global and local explanations could improve people's perceived understanding of the model. We consider both the global understanding of the AI model as a whole and the local understanding on the rationales behind predictions. These questions were skipped in the manual condition.
  - **Understanding of model.** I felt that I had a good understanding of how the AI works.
  - **Understanding of prediction.** I felt that I had a good understanding of why the AI identifies a comment to be toxic.

## 5 RESULTS

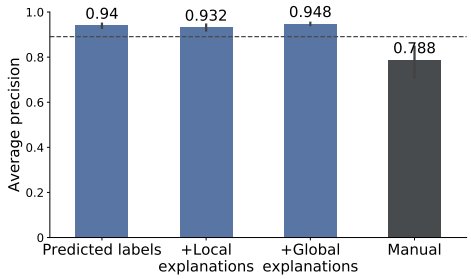
We report results based on the three sets of evaluation measured described above: efficacy, efficiency and engagment, and subjective measures. We refer to the WikiAttack task as in-distribution and Reddit task as out-of-distribution and the terms will be used interchangeably.

### 5.1 Efficacy

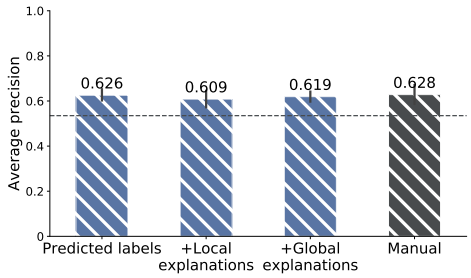
**Even lay people are able to create rules with higher precision than the model working alone, both in-distribution and out-of-distribution (see Fig. 10 and Fig. 11).** To determine whether participants are able to create rules that improve model precision, we conduct $t$-test on the precision of conditional delegation with human-created rules vs. the model working alone. We find that differences are all statistically significant ($p$ <0.001), both on WikiAttack (in-distribution) and Reddit (out-of-distribution), based on average precision and union precision. In particular, on WikiAttack, the model working alone already outperforms the manual condition, and conditional delegation further improves the precision. These observations demonstrate that humans, in our case turkers who are not experts in content moderation, are able to create rules that improve model precision, suggesting that conditional delegation can be a promising direction to pursue.

Next, we examine the effect of distribution types and experimental conditions on precision. We conduct two-way ANOVA of distribution types and experimental condition in average precision and union precision. We find significant effects in distribution type, experimental condition, and their interaction ($p$ <0.001). The effect of distribution type is the most salient, suggesting a clear difference between in-distribution and out-of-distribution.

Given the significant interaction, we further conduct one-way ANOVA to understand the effect of experimental condition on performance separately for WikiAttack and Reddit, and if significant, conduct post-hoc analysis using Tukey's HSD. For average precision (Fig. 10), experimental condition has a significant effect in WikiAttack ($p$ <0.001), but not in Reddit ($p$ =0.864). Post-hoc Tukey's HSD shows that the manual condition is significantly worse than all other experimental conditions with delegation support features ($p$ <0.001) on WikiAttack. For union precision (Fig. 11a and 11b) (using rules created by a participant as a set), experimental condition significantly affects performance in both WikiAttack and Reddit ($p$ <0.001). Post-hoc Tukey's HSD shows on WikiAttack, the




(a) WikiAttack



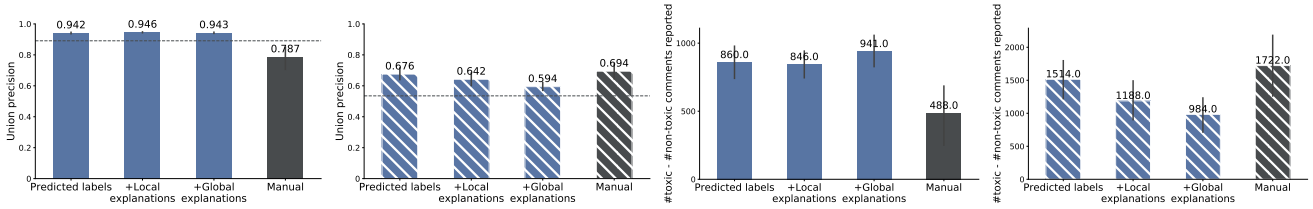

(b) Reddit

**Figure 10: Average precision for WikiAttack (in-distribution) and Reddit (out-of-distribution). Error bar shows 95% confidence interval throughout the paper, and the dashed lines show the precision with the model working alone. The first three conditions represent the precision with conditional delegation, while the manual condition reports precision via the manual rule-based approach by reporting all comments that contain any keyword.**

manual condition is significantly worse than other experimental conditions with delegation support features ($p$ < 0.001). On Reddit, we found the global explanation condition is worse than the manual condition ($p$ =0.004), and only showing prediction labels ($p$ =0.028).

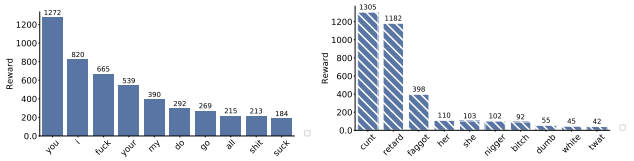
Finally, we examine the effect of distribution types and experimental conditions on reward. Two-way ANOVA finds a statistically significant effect of distribution type and interaction between distribution type and experimental condition ($p$ <0.001). Therefore, we conduct one-way ANOVA to understand the effect of experimental condition on reward separately for WikiAttack and Reddit. On WikiAttack (Fig. 11c), we find a statistically significant effect of experimental condition ($p$ =0.01), and post-hoc Tukey's HSD shows that the manual condition is significantly worse than other experimental conditions with delegation support features ($p$ <0.001 for global explanations, $p$ =0.007 for local explanations, and $p$ =0.004 for predicted labels). On Reddit, the experimental condition also has a statistically significant effect ($p$ =0.018). Post-hoc Tukey's HSD shows that only the difference between the manual condition and global explanations is significant ($p$ =0.018).

These results show that, on WikiAttack, where the model performs well, people can easily identify rules with both high precision (average precision) and with high coverage (reflected by union precision and reward), as long as predicted labels are provided, achieving complementary performance. But on Reddit, where the model's base performance is significantly worse, it is more challenging to achieve better human-AI performance over the manual rule-based



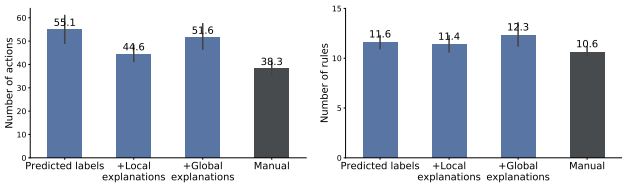

(a) WikiAttack (b) Reddit (c) WikiAttack (d) Reddit

**Figure 11: Union precision and reward for WikiAttack (in-distribution) and Reddit (out-of-distribution). The dashed lines in Fig. 11a and 11b show the precision with the model working alone. Reward is defined as (number of reported toxic comments - number of reported non-toxic comments).**




(a) WikiAttack (b) Reddit

**Figure 12: Top 10 human-created rules in reward when used for conditional delegation.**




(a) Number of actions. (b) Number of rules.

**Figure 13: Engagement. Conditional delegation with all delegation support features leads to much better engagement and more submitted rules than the manual condition.**

approach. It follows that in both situations, we do not observe that explanations, either local or global, significantly improve the performance of conditional delegation. However, adding the global explanation feature seems to unexpectedly hurt people's ability in choosing rules with both high coverage and high precision, and lead to slightly lower human-AI performance in union precision and reward.

**What rules do people make?** To further make sense of their performance, we dive into the content of the rules provided by participants. Table 3 lists the top rules in each condition along with the percentage of people who chose that rule. On WikiAttack, participants with delegation support features are more likely to choose "fuck" (above 60%), a high-precision rule to identify toxic content in Wikipedia comments as shown in Section 3, than the manual condition (only 36.7%). In comparison, for Reddit, "fuck" is a less precise rule (i.e., people also use the word in non-toxic comments). The word does not show up in top 10 for the manual condition, but shows up in the other conditions.

This observation suggests that the delegation support features can help users identify good rules when the model performs well, but may mislead people when the model performs poorly. The reason that global explanations slightly hurt the performance in union precision and reward could be that participants were led to choose some high-coverage rules with relatively low precision such as "fuck", which was listed in the global keywords (Figure 8).

Fig. 12 further shows the top words in reward among the rules created by participants. In addition to "fuck" on WikiAttack and "cunt"/"retard" on Reddit, the result highlights the advantage of conditional delegation. Users can achieve high reward by trusting the model beyond swearing words, for instance, "you" on WikiAttack and "her" on Reddit. The reason is that the AI model excels at deciding whether "you" is used for personal attack or simply for referring purposes.

**Summary.** In short, our results demonstrate that conditional delegation is more effective than the model working alone, and that even laypeople are able to create high-quality conditional delegation rules for content moderation. Compared to the manual rule-based approach currently used for content moderation, advantage of our human-AI collaborative approach via conditional delegation may depend on the base performance of the AI, and may not be sufficient if the AI significantly under-performs, e.g., when used on out-of-distribution comments. Further research is required to understand the necessary conditions for conditional delegation to outperform the manual rule-based approach. Our analysis did not find evidence that model explanations could help people create better rules for conditional delegation. We explore their benefits for other aspects of user experience later.

## 5.2 Efficiency and Engagement

We conduct two-way ANOVA to determine whether distribution type and experimental condition have a significant effect on user engagement (number of actions, number of rules) and efficiency (elapsed time, rules per minute). In all evaluation measures of engagement and efficiency, we only find statistically significant effects of experiment conditions, suggesting that patterns with two distribution types are comparable. Therefore, in this section, we merge the data on WikiAttack and Reddit, and report results from one-way ANOVA on experimental conditions.

**Participants working on conditional delegation are more engaged (see Fig. 13).** Fig. 13a shows that participants with all delegation support features were much more engaged than the manual condition. In particular, predicted labels only condition incurred many more actions than other conditions. One-way ANOVA also finds a statistically significant effect in experimental condition ($p < 0.001$). Post-hoc Tukey's HSD shows the negative difference between the manual condition and other experimental conditions are all statistically significant ($p$ <0.001 for predicted labels and global explanations, $p$ =0.009 for local explanations).

| | |
|---|---|
| | WikiAttack |
| Predicted labels | bitch (69.0%), asshole (62.1%), fuck (62.1%), cunt (62.1%), nigger (51.7%), dick (48.3%), faggot (44.8%), shit (44.8%), fag (37.9%), motherfucker (31.0%) |
| + Local explanations | bitch (71.4%), cunt (71.4%), asshole (67.9%), fuck (60.7%), faggot (53.6%), pussy (42.9%), dick (39.3%), fag (35.7%), cock (35.7%), retard (35.7%) |
| + Global explanations | faggot (86.7%), nigger (73.3%), fuck (70.0%), bitch (66.7%), cunt (56.7%), cock (56.7%), ass (50.0%), asshole (46.7%), shit (46.7%), pussy (36.7%) |
| Manual | cunt (70.0%), nigger (63.3%), faggot (60.0%), fag (56.7%), bitch (53.3%), asshole (46.7%), retard (43.3%), whore (43.3%), fuck (36.7%), pussy (26.7%) |
| | Reddit |
| Predicted labels | cunt (86.2%), bitch (72.4%), faggot (62.1%), fuck (58.6%), retard (44.8%), asshole (41.4%), nigger (41.4%), pussy (34.5%), fag (31.0%), whore (31.0%) |
| + Local explanations | cunt (72.4%), bitch (65.5%), fuck (55.2%), pussy (55.2%), nigger (48.3%), asshole (41.4%), faggot (41.4%), retard (37.9%), shit (37.9%), dumbass (34.5%) |
| + Global explanations | bitch (69.0%), cunt (65.5%), faggot (62.1%), fuck (58.6%), retard (58.6%), nigger (41.4%), pussy (41.4%), shit (37.9%), dick (37.9%), idiot (34.5%) |
| Manual | nigger (76.7%), cunt (73.3%), faggot (60.0%), bitch (56.7%), retard (43.3%), whore (43.3%), asshole (36.7%), fag (30.0%), spic (30.0%), chink (30.0%) |

**Table 3: Most frequent rules chosen by participants.**

When it comes to number of rules, the outcome of task engagement, the difference is not as salient. Because we require a minimum of 10 rules, every condition leads to a little above 10 rules: the manual condition is just above 10 at 10.6, while global explanations leads to 12.3 rules. That said, one-way ANOVA still finds a significant effect in experimental condition ($p$ =0.021). Post-hoc Tukey's HSD shows that the difference between global explanations and the manual condition is statistically significant ($p$ =0.028).

**Explanations improve task efficiency (see Fig. 14a and 14b).** Fig. 14a shows the time spent on the interactive interface in each condition: participants with predicted labels only spent the most time on this task. This result is consistent with the number of actions because it likely requires more time to take more actions. However, the difference is relatively weak: one-way ANOVA shows that the effect of experimental conditions is only borderline significant ($p$ =0.074), and post-hoc Tukey's HSD do not find any statistically different pairs.

Rules per minute is a better measure of efficiency since it reflects how long it takes for people to identify a rule that they are satisfied with. The trend is reversed from the time spent: predicted labels only lead to the lowest number of rules per condition, however, with the help of explanations, humans are able to achieve a greater number of rules per minute. One-way ANOVA confirms a statistically significant effect of experimental condition ($p$ =0.008). Post-hoc Tukey's HSD suggests that the only statistically different pair is predicted labels only and global explanations ($p$ =0.031), suggesting that global explanations helped participants to achieve the highest efficiency to come up with rules.

**Global explanations lead to much higher overlap between most frequent words in rationales (Fig. 14c).** To understand this improvement in efficiency, we examine the overlap between human-created rules and the most frequent words in rationales (Table 2), which are the words shown in global explanations and also more likely to have appeared in the highlighted words in local explanations. Fig. 14c shows that global explanations lead to much higher overlap than the other conditions. This observation confirms that global explanations provide direct hints for possible rules, thus improved the efficiency to come up with required number of rules.

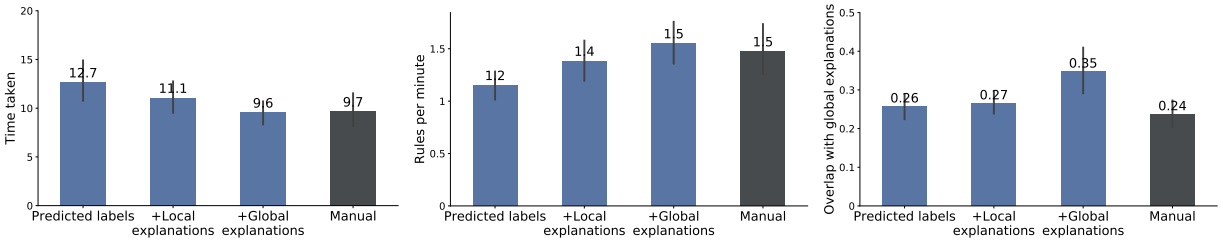
**Summary.** Taken together, our results show that people are more engaged when performing conditional delegation than working on creating manual rules, with more actions and a tendency to spend more time on the task. This tendency comes with a cost of efficiency in creating rules when only showing predicted labels. Showing model explanations, especially global explanations, can significantly improve the efficiency, resulting in comparable efficiency between conditional delegation and the manual rule-based approach. The reason can be attributed to participants leveraging keywords in explanations as hints to create delegation rules. Future research is required to explore means to encourage people to examine the performance of these hinted rules more carefully, to improve both efficiency and efficacy.

## 5.3 Subjective Perception

Finally, we report the subjective perception of participants (subjective workload, confidence, and perceived understanding of AI). All results are based on answers in exit survey, with a five-point Likert scale.

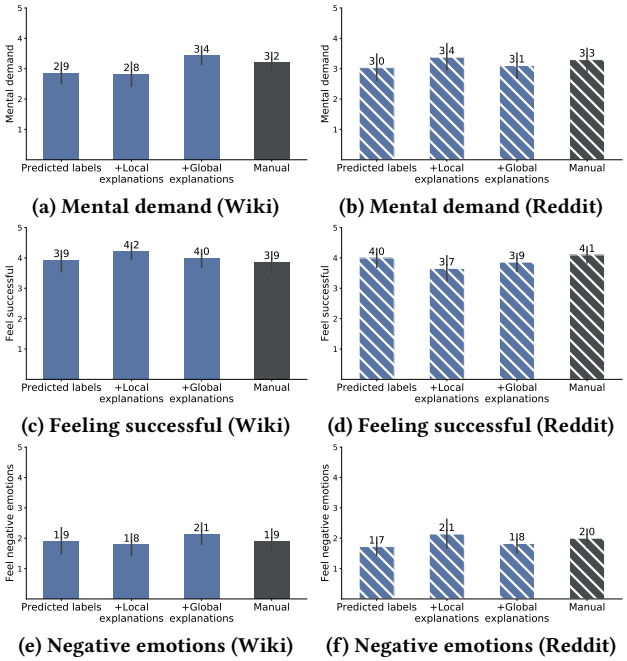
**Subjective workload.** Overall, participants were neutral about whether the task was mentally demanding (M=3.15, SD=1.08), agreed that they felt relatively successful in accomplishing the task (M=3.95, SD=0.91), and disagreed that they felt negative emotions (M=1.93, SD=1.03). Two-way ANOVA does not show any statistically significant effects of distribution types and experimental conditions. For WikiAttack, we observe a weak trend that local explanations lead to less subjective workload (lower mental demand, more feeling of success, and less negative emotion) while adding global explanation has the opposite effect. These patterns, however, do not hold for Reddit.

**Confidence.** Overall, participants reported relatively strong confidence in all of our measures: confidence in the model (M=3.63, SD=1.01), confidence in the rules they created (M=3.73, SD=0.98), confidence in the deployment of the system from human-AI collaboration (M=3.61, SD=1.0), leaning towards agreeing with all these




(a) Elapsed time (b) Rules per minute (c) Overlap with global explanations.

**Figure 14: Efficiency. Participants spent more time working on conditional delegation than the manual condition, but the efficiency is improved with explanations, especially global explanations.**

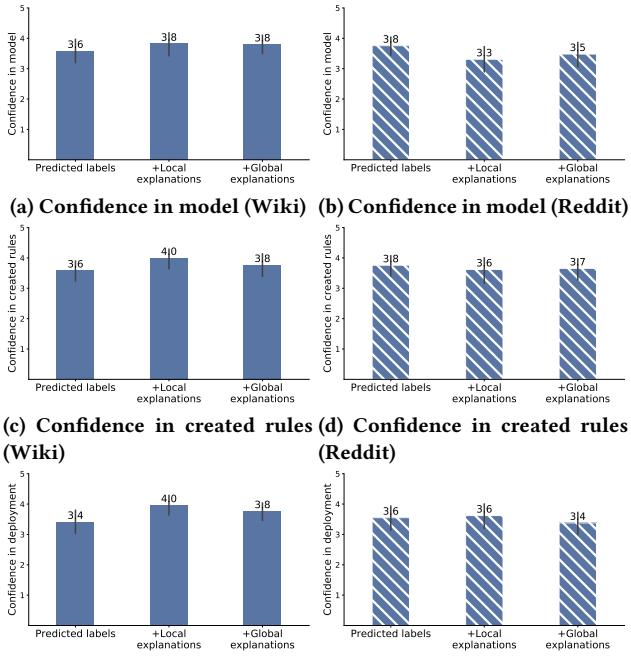



(a) Mental demand (Wiki) (b) Mental demand (Reddit)

(c) Feeling successful (Wiki) (d) Feeling successful (Reddit)

(e) Negative emotions (Wiki) (f) Negative emotions (Reddit)

**Figure 15: Subjective workload. Overall, participants were neutral about whether the task was mentally demanding (M=3.15, SD=1.08), agreed that they felt successful in accomplishing the task (M=3.95, SD=0.91), and disagreed that they felt negative emotions (M=1.93, SD=1.03).**




(a) Confidence in model (Wiki) (b) Confidence in model (Reddit)

(c) Confidence in created rules (Wiki) (d) Confidence in created rules (Reddit)

(e) Confidence in deployment (Wiki) (f) Confidence in deployment (Reddit)

**Figure 16: Confidence. Overall, participants show strong confidence in the model, their performance, and moderators' potential adoption.**

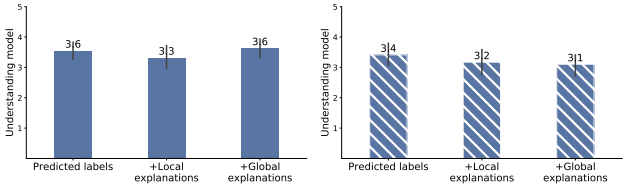
statements. We do not find any statistically significant effect of distribution type and experimental condition with two-way ANOVA. There is a non-significant trend that conditions with explanation, especially local explanation, result in better confidence for WikiAttack, but not for Reddit, where the model performs relatively poorly.

**Perceived understanding.** Overall, participants report a good global understanding of the model (M=3.37, SD=0.97) and local understanding of individual predictions (M=3.56, SD=1.05) on WikiAttack than Reddit, possibly related to the difference in model performance between distribution types. Two-way ANOVA only shows a marginally significant effect of distribution type in global understanding of the model ($p$ =0.063). It is somewhat surprising that model performance leads to this difference in perceived understanding. Interestingly, there is a trend that local explanations lead to better perceived local understanding on predictions, but worse perceived global understanding on the model, but not when global explanations are added.

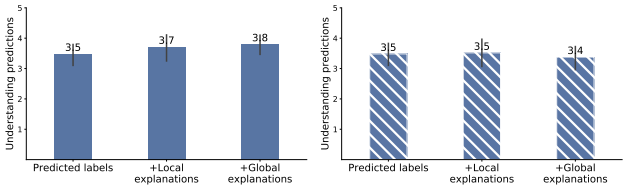
**Summary.** Overall, subjective measures show a relatively positive experience across board, but not strong difference between conditions. There is some evidence that when the model performs well (WikiAttack), local explanations provide the best experience: strong performance with relatively high efficiency, less subjective workload and more confidence in the outcomes.

## 6 DISCUSSION

Through investigating the three research questions introduced in the beginning, our study shows the promise of conditional delegation as a new paradigm for human-AI collaboration. Even with crowdworkers who are not experts of the content moderation task, conditional delegation can achieve better performance than the




(a) Understanding model (Wiki) (b) Understanding model (Reddit)




(c) Understanding predictions (Wiki) (d) Understanding predictions (Reddit)

**Figure 17: Understanding. Overall, participants report a better understanding of the model as a whole and individual predictions on WikiAttack than Reddit, although the differences are not statistically significant.**

model working alone. However, whether the human-AI collaboration can outperform the manual rule-based approach varies for in-distribution and out-of-distribution AI. Out-of-distribution AI has significant performance disadvantage that cannot be adequately compensated by conditional delegation. We also found that, in general, providing predicted labels with our keyword search based interface is sufficiently effective in supporting people to create delegation rules. Providing explanations can improve efficiency by hinting on rules to consider, but can also mislead people to use high-frequency but not necessarily high-precision rules. We discuss implications of these results below.

**The promise of conditional delegation.** Our study is a first step towards understanding and leveraging the promise of conditional delegation. It is an intuitive approach that can be used in a wide variety of domains so that users can proactively decide when to use an AI model and in what ways based on the output of the AI. For instance, judges can specify when to show the risk estimates for recidivism prediction and when to hide the model output. Doctors can identify subsets of patients for which they rely on AI to send alerts. Our study only explored one type of workflow and one type of action. Different applications may require a diverse set of workflows and actions, and have varying tradeoffs between false positives and false negatives.

Moreover, we only begin to define the design space for supporting users in conditional delegation. An essential requirement is to help users make sense of model behaviors under different delegation conditions. Keyword-based rules are a reasonable approach in content moderation given that rule-based methods are already used in AutoModerator. We used a rationale-style model to facilitate this kind of interaction, although we expect post-hoc explanation methods to play similar roles. It is important to empirically validate the effect of the underlying models and explanation methods. We also focused exclusively on conditional delegation based on trustworthy regions in the input space. A promising direction is to investigate the joint effect of delegation based on inputs and outputs [18, 47]. Another limitation of our study lies in that crowdworkers only came up with about 10 rules because of our minimum requirement. Although our results are encouraging with non-expert users, additional experiments are required to validate the potential of conditional delegation with expert users. Notably, future research can explore means to facilitate people to identify a greater number of rules, examine the combined effect of rules, and monitor the performance of rules after model deployment. In long-term deployment, it is especially valuable to investigate how to update the delegation conditions once the model is updated.

Additionally, conditional delegation can potentially alleviate AI bias as we give users the freedom to choose trustworthy regions based on their domain knowledge or notion of fairness. However, the flip side is that this process could introduce human bias, if for example one's notion of fairness is ill conceived. We encourage future work to understand and develop ways to rail-guard the impact of human biases in conditional delegation.

**The effect of distribution shift.** Our results highlight the importance of considering the effect of distribution shift in designing experiments on human-AI collaboration, to better understand the generalizability of results. We are able to achieve complementary performance in-distribution on WikiAttack, but not out-of-distribution on Reddit. In practice, it is rarely the case that an AI model faces exactly the same distribution in deployment. Therefore, it is critical to understand the outcomes in out-of-distribution contexts to understand the generalizability or applicable scope of a given form of human-AI collaboration.

It is useful to note that although our results are presented as in-distribution vs out-of-distribution, the differences are complicated between WikiAttack and Reddit. First, there exists a clear difference in model performance, so our results can be seen as comparing a high-performance model with a low-performance model. Second, the nature of comments on Wikipedia and Reddit differs substantially. It is possible that crowdworkers are more used to comments on Reddit or that common swearing words such as "retard" and "cunt" happened to work well on the Reddit dataset that we used. This complexity demonstrates that the contrast of in-distribution versus out-of-distribution contexts is not a monolithic dimension, which further adds to the challenge of experimental design to account for the effect of distribution shift.

**The priming effect of explanations.** While explanations can improve efficiency, global explanations are found to slightly hurt performance when working with out-of-distribution AI, as participants may have chosen the keywords in explanations without carefully examining the model behaviors with them. These observations echo concerns of unintended consequences with the use of explanations in human-AI collaboration [8, 37, 51].

In other words, for our task of creating keywords rules, keywords-based explanations have a priming effect that leads to biased adoption of presented words. Note that priming, if used appropriately, can shape user behaviors in a positive way. The challenge is that with the technique we used to generate global explanations (i.e., most frequent tokens in rationale), the top tokens do not necessarily correlate with high precision (Fig. 6). Future work can explore techniques that can exploit some proxy of precision, such as considering the uncertainty or confidence of predictions. Another direction is to utilize de-biasing technique to mitigate the effect of priming,

such as explicitly reminding people to attend to wrong predictions with the chosen keywords.

It is worth noting that local explanations seem to have less of a priming effect than global explanations but still improves efficiency. It is possible that the many highlights in search results are too scattered to have a salient effect. Future work can explore other XAI techniques or provide additional support, such as to help users have an overview of the rationales in all search results.

**Implications on content moderation.** It is impressive that crowdworkers can already create keyword-based rules that achieve greater precision than the model working alone. However, we recognize that our experiment setup is only a first step towards using conditional delegation in content moderation. First, crowdworkers are not representative of moderators, who have way more experience with their platform's data. As moderators are more familiar with the moderation process and more knowledgeable about important words, experts might find the interface more useful than crowdworkers. However, participatory design and future work can develop more serendipitous features. Second, in practice, moderators usually have historical data on which moderation decisions were made. This historical data can be used in the process of creating keyword-based rules. Third, prior work has shown that moderators often update the rules used by AutoModerator [18, 44] and our work does not take into account any future updates. Neither do we leverage any existing rules that moderators have created. For future work, we hope to integrate a model that receives feedback from moderators and allow updates to the model to reflect the feedback. The ideal pipeline would require careful development in the model architecture and interface to refrain any unnecessary actions from interfering with moderators' tasks. Last but not least, content moderation involves a wide range of different rules beyond toxicity, and even the policies under the umbrella term of toxicity can vary, so the AI model that we uses represents a narrow component in content moderation. In short, our work uses content moderation as a testbed to illustrate the promise of conditional delegation. Much future work is required to realize the impact of conditional delegation in content moderation.

**Limitations.** First, our work represents one instantiation of conditional delegation. We emphasize precision and coverage to increase the ability of moderators to deal with a large amount of comments ("true positives") while minimizing unnecessary labor for moderators ("false positives"). This tradeoff between true positives, true negatives, false positives, and false negatives can vary in practice depending on the application and the actions taken according to AI predictions. Second, our participants are not representative of content moderators. It also follows that our evaluations are limited by the number of rules that participants created in about 10 minutes. Our case study shows the promise of conditional delegation, but further study is required in each application domain of interest to develop the best design for human-AI collaboration in identifying delegation conditions. Third, our choice of model, datasets, and explanations affect the experimental outcome. It is important to further dissect the relevant dimensions and investigate the effect of alternative choices.

## ACKNOWLEDGMENTS

We thank all anonymous reviewers for their insightful suggestions and comments. We thank all members of the Chicago Human+AI Lab for feedbacks on early versions of our website interface. All experiments were approved by the University of Colorado IRB (21-0385). This work was supported in part by NSF grants IIS-1837986, 2125116, and 2125113.

## REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

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

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[4] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2020. OpenCrowd: A Human-AI Collaborative Approach for Finding Social Influencers via Open-Ended Answers Aggregation. In *Proceedings of The Web Conference 2020*. 1851–1862.

[5] Zahra Ashktorab, Q Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–20.

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[8] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[10] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.

[11] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365, 6456 (2019), 885–890.

[12] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[13] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[14] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 4.

[15] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. Hello AI: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 104.

[16] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[17] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of CSCW* (San

Francisco, California, USA) *(CSCW '16)*. ACM, New York, NY, USA, 1201–1213.

[18] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[19] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *In Proceedings of ACM Human Computer Interaction* 1, Computer-Supported Cooperative Work and Social Computing, Article 31 (2017), 22 pages.

[20] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *In Proceedings of ACM Human Computer Interaction* 2, Computer-Supported Cooperative Work and Social Computing, Article 32 (Nov. 2018), 25 pages.

[21] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of WWW*. 184–195.

[22] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 559.

[23] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[24] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4069–4082.

[25] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2382–2393.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

[27] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.

[28] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[29] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.

[30] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[31] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! Characterizing an Ecosystem of Governance. In *Proceedings of International AAAI Conference on Web and Social Media(ICWSM)*.

[32] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2020. Explainable Active Learning (XAL): An Empirical Study of How Local Explanations Impact Annotator Experience. *arXiv preprint arXiv:2001.09219* (2020).

[33] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven.

[34] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[35] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[36] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 90–99.

[37] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 50.

[38] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[39] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of ICCV*.

[41] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of CSCW*.

[42] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[43] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[44] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[45] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[46] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2021–2031.

[47] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2021. Towards Unbiased and Accurate Deferral to Multiple Experts. *arXiv preprint arXiv:2102.13004* (2021).

[48] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* (2012), 125–178.

[49] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[50] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[51] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[52] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).

[53] 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).

[54] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[55] Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. 2020. The limits of human predictions of recidivism. *Science advances* 6, 7 (2020), eaaz0652.

[56] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *arXiv preprint arXiv:2101.05303* (2021).

[57] Brian Lubars and Chenhao Tan. 2019. Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability. In *Proceedings of NeurIPS*.

[58] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.

[59] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[60] J Nathan Matias. 2016. The civic labor of online moderators. In *Internet Politics and Policy conference. Oxford, United Kingdom*.

[61] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3428–3448.

[62] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[63] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 189–199.

[64] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[65] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[66] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2020. Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors* (2020), 0018720820960865.

[67] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.

[68] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2799–2804.

[69] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[70] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4755–4764.

[71] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[72] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign.

[73] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of ACL*.

[74] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A Framework of Severity for Harmful Content Online. *arXiv preprint arXiv:2108.04401* (2021).

[75] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In *Proceedings of CSCW*. ACM, New York, NY, USA.

[76] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[77] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[78] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community. In *Proceedings of CSCW*.

[79] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies* 67, 8 (2009), 639–662.

[80] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[81] Dakuo Wang, Q Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. 2021. How Much Automation Does a Data Scientist Want? *arXiv preprint arXiv:2101.03970* (2021).

[82] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[83] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[84] Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. *arXiv preprint arXiv:1907.03324* (2019).

[85] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of NAACL*.

[86] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. *arXiv preprint arXiv:2005.00582* (2020).

[87] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[88] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399.

[89] 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 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[90] Yiming Zheng, Serena Booth, Julie Shah, and Yilun Zhou. 2021. The Irrationality of Neural Rationale Models. *arXiv:2110.07550 [cs]* (Oct. 2021). http://arxiv.org/abs/2110.07550 arXiv: 2110.07550.