# Using the Crowd to Prevent Harmful AI Behavior

TRAVIS MANDEL, JAHNU BEST, RANDALL H. TANAKA, HIRAM TEMPLE, CHANSEN HAILI, and SEBASTIAN J. CARTER, University of Hawai'i at Hilo, USA
KAYLA SCHLECHTINGER, University of Minnesota, USA
ROY SZETO, Center for Game Science, USA

To prevent harmful AI behavior, people need to specify constraints that forbid undesirable actions. Unfortunately, this is a complex task, since writing rules that distinguish harmful from non-harmful actions tends to be quite difficult in real-world situations. Therefore, such decisions have historically been made by a small group of powerful AI companies and developers, with limited community input. In this paper, we study how to enable a crowd of non-AI experts to work together to communicate high-quality, reliable constraints to AI systems. We first focus on understanding how humans reason about temporal dynamics in the context of AI behavior, finding through experiments on a novel game-based testbed that participants tend to adopt a long-term notion of harm, even in uncertain situations that do not affect them directly. Building off of this insight, we explore task design for long-term constraint specification, developing new filtering approaches and new methods of promoting user reflection. Next, we develop a novel rule-based interface which allows people to craft rules in an accessible fashion without programming knowledge. We test our approaches on a real-world AI problem in the domain of education, and find that our new filtering mechanisms and interfaces significantly improve constraint quality and human efficiency. We also demonstrate how these systems can be applied to other real-world AI problems (e.g. in social networks).

CCS Concepts: • **Human-centered computing** → *Interaction paradigms*; **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence.*

Additional Key Words and Phrases: artificial intelligence, constraints, human-AI collaboration, data quality, crowdsourcing

**ACM Reference Format:**
Travis Mandel, Jahnu Best, Randall H. Tanaka, Hiram Temple, Chansen Haili, Sebastian J. Carter, Kayla Schlechtinger, and Roy Szeto. 2020. Using the Crowd to Prevent Harmful AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 97 (October 2020), 25 pages. https://doi.org/10.1145/3415168

## 1 INTRODUCTION

As AI systems are given greater agency to impact human lives, we must ensure that these systems do not engage in harmful behavior. Although preventing physical harm (e.g. an autonomous car crash) has received considerable attention recently, non-physical harm is also a topic of great

Authors' addresses: Travis Mandel, tmandel@hawaii.edu; Jahnu Best, jahnub@hawaii.edu; Randall H. Tanaka, dh404@hawaii.edu; Hiram Temple, htemple@hawaii.edu; Chansen Haili, haili808@hawaii.edu; Sebastian J. Carter, sjc7@hawaii.edu, University of Hawai'i at Hilo, 200 W. Kawili St, Hilo, Hawai'i, USA, 96720; Kayla Schlechtinger, schle543@umn.edu, University of Minnesota, 200 Union Street SE, Minneapolis, Minnesota, USA, 55455; Roy Szeto, roylszeto@gmail.com, Center for Game Science, Box 352350, Seattle, Washington, USA, 98195-2350.

2573-0142/2020/10-ART97 $15.00
https://doi.org/10.1145/3415168

concern.[1] One recent example of this is online movie recommendations [5], where we do not want an AI system to recommend a movie with extreme violence to a young child, even if they are likely to watch it. Similar problems arise in education, where we do not want to allow a mathematical AI tutoring system to ever teach incorrect or misleading information to young students.[2]

A standard machine learning approach to this problem is simply to learn through trial-and-error which actions are harmful. However, actually trying a potentially harmful action in the real system is extremely undesirable in most settings. Indeed, numerous communities have acknowledged the insufficiency of systems that learn about harmful actions only after making a (potentially harmful) decision [2], and this approach could also lead to numerous legal issues [20]. Therefore, researchers have studied AI systems that learn to predict when actions are harmful before they are executed, for instance based on trying them in a simulator [24]. Unfortunately, this requires a near-perfect simulator to ensure safe behavior, which is unachievable in most real-world applications. Therefore, such systems often rely on experts (in both AI as well as the specific domain) to hand-craft rules, called *constraints*, that specify what behaviors are harmful. In this setting, AI researchers tend to focus on studying how to create machine learning systems that obey these prespecified constraints [13, 16, 54].

Although this research is valuable, often the "weakest link" in using behavioral constraints are the human factors involved in specifying them in a manner that correctly captures our human understanding of harm. Perhaps the most influential work in this space is the Moral Machine project by Awad et al. [4], which studies how to understand human preferences surrounding the safety of autonomous cars.

While undoubtedly impactful, the Moral Machine project is not without its limitations. It focuses on simplified fictional scenarios (e.g. killing passengers or killing pedestrians), which have been broadly criticized as far removed from the scenarios an autonomous car would likely encounter in the real world [7, 9]. In their response to these criticisms, Awad et al. [3] states that the purpose of the simplification is to tease out general-purpose ethics rules to guide autonomous car development. Yet, as others have pointed out [22], the manner in which this highly complex real-world problem is simplified itself introduces a considerable bias.[3] This bias is very concerning, as it may further shift power to a small group of already-powerful companies and executives [31], while simultaneously allowing them to claim that "society" is involved (albeit in a token fashion). Although it may not be the case that AI developers and tech companies are intentionally trying to marginalize the opinions of a larger and more diverse group of stakeholders, such marginalization can easily occur subconsciously [6].

In order to allow individuals with a diverse set of backgrounds to participate in real-world decisions surrounding AI harm, we must first ensure that the AI harm problem is posed to users in a way that is both accessible to non-AI experts, and minimizes the bias introduced by the AI designer. The latter goal is in line with work by Constanza-Chock in Design Justice [14] which suggests the role of the designer should be minimized, as they should serve as a facilitator who

[1] Various communities make distinctions between the broader term "harm" and the narrower term "safety". For instance, in the AI community the term "safety" is usually used only to refer to physical harm, hence we use the broader term "harm" in our paper. But it is important to note that the scope of the paper is focused only on "harms" that arise directly as the result of specific AI actions (e.g. an AI saying something offensive), and does not address more abstract harms that are inherent to the system design itself (such as the harm of a self-driving truck putting professional truck drivers out of work). Here we assume that the general task we are trying to automate has been found to be sound with respect to harm, and the focus is on preventing the AI system from taking certain actions in that task that could be considered harmful.

[2] In other domains, we may have to allow a small possibility of teaching incorrect information due to the fact that the field is controversial or constantly evolving (for instance, our understanding of nutrition has changed considerably over the last several decades [33]). Mathematics does not suffer from this issue.

[3] For example, by eliminating uncertainty from the scenarios, researchers eliminate a key element of human behavior [48].

helps center the voices of community members rather than an expert who prioritizes their own agenda. However, while a Design Justice philosophy argues that all members of a community should be viewed as *"expert[s] based on their own lived experience"* and thus given an equal voice in the process [14], when faced with complex real-world AI situations, past AI work has questioned whether people without any formal expertise are capable of judging the long-term safety of real-world AI behaviors [3, 50]. This is a legitimate concern, as defining constraints in complex real-world environments can be quite challenging, and human mistakes or misunderstandings can have serious consequences. A constraint that is too permissive could result in extremely harmful behavior, while one that is too strict may result in an AI system that is incapable of effectively performing its assigned task. Therefore, in this paper, we present the first study of how to make the process of specifying AI constraints more accessible to non-AI experts, with a focus on presenting users with randomly-chosen real-world situations instead of ones handcrafted by a designer.

Despite acknowledging that simplifications relating to AI harm or safety may themselves be harmful, at the same time we must acknowledge that simplifications have an important (albeit limited) place in studying how individuals reason about AI safety. Indeed, behavioral fields such as psychology and economics have a long history of designing simplified experiments to better understand principles of human behavior [40, 49]. The same is true in AI safety: for instance, a study was carried out which, by asking certain simplified questions, revealed the valuable insight that human perception of the morality of AI behavior depends greatly on their perception of the applicable laws [29]. As such, the problem is not simplification itself, but rather using the simplified decisions as a direct substitute for decisions people would make in the real world, without fully considering all the nuances of the real problem.

Therefore, we take a two-pronged approach in this paper. First, we design a simplified experiment to elucidate novel questions surrounding how crowd workers reason about the long-term effects of AI harm when different groups are impacted. Although this gives us some insight into the ability of non-AI experts to reason long-term about safety, even when it affects others, we acknowledge that this initial study is limited by the fact that it is artificial. Therefore, we then turn our attention to studying how people make AI harm decisions in a complex real-world problem in the domain of education. In this context, we identify new techniques that boost constraint quality through better filtering and promoting worker reflection. We also develop a novel rule-based approach, which makes the process of writing rules accessible to non-AI experts, and utilizes worker effort much more efficiently while improving constraint quality.

## 2 RELATED WORK

### Preventing AI Harm

Saunders et al. [50, 51] studies how humans can prevent AI harm through blocking undesirable actions in real time. However, their experiments were conducted using standard Atari videogames as testbeds; therefore, the authors artificially invented the meaning of "harm" in these cases. Additionally, the only participants in their experiments were two authors of the paper, and as such, they did not examine the quality of constraint information provided by humans. Saunders et al. leaves open the question of how effective humans are at avoiding long-term harm (i.e. *non-local catastrophes*) and frames how to increase human efficiency when communicating harm to AI systems as an explicit open question. We carefully study both issues in this paper.

Recent work by Thomas et al. [55] has examined methods to prevent undesirable AI behavior, which essentially requires people to specify only harmful outcomes instead of harmful actions. Unfortunately, this approach inevitably results in the possibility that undesirable outcomes occur as the system is learning the effects of new actions, and hence is not suitable for outcomes that are harmful enough that we want to ensure they will never occur.

**Specifying constraints** Although specifying behavioral constraints is well-known to be a important prerequisite to building effective AI systems, relatively little work has studied the process by which this occurs. Traditionally, experts (in both AI and the domain) write constraint rules in a specialized programming language such as LISP [39] or EQL [12]. An alternate approach is to use machine learning to learn constraints from safe and unsafe examples [5]; however, this is undesirable as if the machine learning system makes an inaccurate prediction, harm can result.

Work by Zhuo [60] (and Gao et al. [26]) looks at a related problem, in which they ask users to help specify an action model (including constraints) for various simulated AI planning tasks (for example, Blocks World). However, instead of having users construct their own constraints, users were simply asked to determine whether a given constraint was correct. Also, this work did not investigate how to improve human performance in specifying constraints.

**Data quality** Numerous methods have been developed to increase data quality in crowdsourcing settings [11, 23, 25, 36, 41]. However, these works do not consider preventing harmful AI behavior, and thus cannot examine or leverage any specific properties of constraint specification tasks (e.g. how to weight precision over recall, the relationship between state and action, etc.). We examine and extend several of these methods in our experiments.

It is standard practice to use gold questions as a means to filter crowd workers [8, 15], but this typically requires substantial expert effort to annotate the gold dataset. Some work has studied programmatically generating gold questions [43], however it still requires experts to spend significant time identifying common mistakes and construct mutation operators and associated error descriptions. In this paper, we propose filtering methods for AI constraints that do not require additional expert annotation effort.

**Crowdsourcing complex work** Much CSCW work has studied how tasks, once reserved for experts, can be made accessible to a crowd of non-expert workers. For instance: Lee et al. [34] studies how crowd workers can best work with experts in real-time to collaboratively sketch prototypes, Li et al. [36] studies how crowd workers can best contribute to the complex process of sensemaking, and van Berkel et al. [56] studies how workers can best reach agreement on notions of fairness, which can often be quite complex [42]. We add to this conversation, studying how to best enable non-AI experts to contribute to the important but complex problem of AI harm.

**Human-AI interaction** The CSCW community has also devoted much study to the problem of how to facilitate effective human-AI interaction. Work has shown that even the most enthusiastic adopters of AI systems have key pieces of the task that they feel should not be automated [58]. A potential reason for this is that even if a machine makes a correct decision, its underlying understanding of the task may differ substantially from human understanding [59]. This leads to concerns about whether such systems can be trusted with important tasks. Mackeprang et al. [37] explores this issue, trying to determine the "sweet spot" on the "level of automation" scale [45] where there is enough automation to be helpful, but not so much that the automation becomes harmful. Although valuable, one difficulty with this "level of automation" approach arises in domains such as tutoring systems and autonomous driving, where even though automated decision-making (automation level 7+) could be enormously helpful, it also opens the door to significant potential harm. Therefore, in this paper we take this conversation a step further, studying the important question of how people can most effectively define *where* (i.e. in what situations) the AI can safely control decision-making.

## 3 PROBLEM SETUP

We consider problems in a standard Markov Decision Process (MDP)-style AI framework [53]. Specifically, we assume a set of actions $\mathcal{A}$ and a (possibly large) set of states $\mathcal{S}$. The environment produces a state $s \in \mathcal{S}$, the AI agent chooses an action $a \in \mathcal{A}$, and the environment transitions to a new state $s' \in \mathcal{S}$ with probability given by a (possibly unknown) transition function $T(s, a, s')$. We assume there exists a function $C(s, a)$, which is unknown to the agent, that returns true if and only if the action $a$ is not harmful to take at state $s$. In this paper, our goal is to use humans to find a Boolean function $\hat{C}(s, a)$ that approximates $C(s, a)$ as well as possible, so that we can develop "safe" agents that only take actions where $\hat{C}(s, a) = true$. Other than this, we are agnostic to the details of the AI algorithm that the agent uses to select actions (this could include reinforcement learning approaches, recommendation algorithms, etc.).

In terms of evaluating $\hat{C}(s, a)$, our primary aim is to maximize **precision**; that is, the proportion of cases where $\hat{C}(s, a) = true$ that $C(s, a) = true$. This is because, since we intend to feed $\hat{C}(s, a)$ to an autonomous agent, if we have a false positive ($\hat{C}(s, a) = true$ and $C(s, a) = false$), the agent may take a harmful action. In this case, high recall is not a priority, as we would much prefer to mark many situations unsafe when they are in fact safe from harm (false negative), than mark a single situation safe from harm when it is actually unsafe (false positive). Of course, the main flaw with only considering precision is that one can achieve high precision by only marking a very small number of situations as safe (and having them be correct). Therefore, the number of cases where $\hat{C}(s, a) = true$ is a secondary consideration.[4]

We make the following simplifying assumption:

Assumption 1. *For every action $a'$, we can produce at least one state $s'$ such that $C(s', a') = true$; in other words, a state where $a'$ is known to be safe.*

This assumption is not very strong, since actions are often constructed with a specific situation in mind. For example, a telemarketing AI may have a script giving the ideal situation for each line of dialog. If this is not the case, it should be relatively easy for an expert to identify some state where the action is safe to try.[5]

## 4 LONG-TERM VS IMMEDIATE HARM

Our problem setup, and indeed most other standard formulations [53], assume constraints are specified on pairs of states and actions. However, in most environments, the agent interacts with the environment through a sequence of actions rather than a single action. As such, there are at least 3 possible interpretations of what is meant by a "harmful" action in a certain state:

**Immediate:** $C(s, a) = false$ if and only if unsafe outcomes will occur immediately after $a$ is taken at $s$.

**Next Action:** $C(s, a) = false$ if and only if, after $a$ is taken at $s$, and before the next state $s_m$ is reached where there are multiple possible actions, unsafe outcomes will occur.

**Long Term:** $C(s, a) = false$ if and only if after $a$ is taken at $s$, for all possible future state-action sequences $s, a, s_2, a_2, s_3, a_3, \ldots$ unsafe outcomes will occur.[6]

To illustrate the differences between these three theories, consider a case where we are evaluating the harmfulness of action $a'$ at state $s'$. $a'$ does not result in immediately harmful outcomes, but it leads to a state $s''$ where all actions immediately result in harmful outcomes. For instance, suddenly stopping an autonomous car on a highway may prevent colliding with an obstacle, but seconds later will lead to an inevitable crash with the cars behind it, no matter the next action taken.

[4] The number of cases marked safe (i.e. the number of positive responses) forms a rough proxy for recall in this case, while being much easier to calculate.

[5] If it is hard to identify any states where the action is safe, it is unclear why the agent was provided this action to begin with.

[6] For clarity, we use absolute terms here; in practice humans likely adopt a weaker definition to account for the inherently uncertain nature of many real-world environments. We further explore the notion of uncertainty in our CarefulCar experiments.

If the dynamics of the system are known to the AI, all theories are operationally equivalent. For example, an MDP planner can infer that even though $a'$ is not immediately harmful, there is no possible future sequence of actions and states which does not terminate in a harmful outcome.

However, when the dynamics are unknown to the AI, which is usually the case in real-world systems, different theories result in significant practical differences. For example, consider a AI agent provided with immediate constraints and located at state $s'$. It sees that $a'$ is allowed by the constraints, and (not knowing that $(s', a') \rightarrow s''$) is likely to execute $a'$, leading to harmful behavior once $s''$ is entered. Therefore, if people tend to produce immediate constraints, to ensure safety we would need mechanisms that explicitly ask humans to reason about the harmfulness of future state-action *sequences* instead of simply a current state and action. On the other hand, if people tend to reason about long-term harm even in the absence of further instructions or prompting, these mechanisms are unnecessary, as obeying long-term constraints ensures that the agent never reaches a state where all actions are harmful.

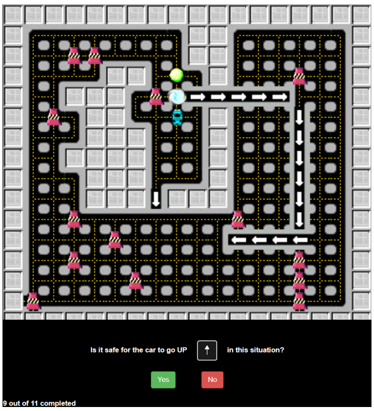
Therefore, better understanding how people reason about long-term harm is key when determining how to break down the complex task of defining constraints on AI systems.

In addition to examining the question of how people reason about long-term harm, we also wish to examine how users' reasoning changes based on **who** is harmed. This is an important question because many biologists subscribe to the notion that evolution leads to organisms that are largely self-interested [57]. Given this, one might expect humans to be fairly proficient at reasoning about harm to themselves, as they are essentially "hard-wired" to do so. But, it might be more difficult to make judgments about how harm impacts others. For instance, it might make a difference whether the harm was to themselves vs. a close friend vs. a random outsider. In the extreme case, where what is being harmed is simply an abstract AI or an autonomous car, it is unclear how well humans will reason about harm to this non-biological entity.[7]

Philosopher Peter Singer provides a theoretical lens [46] through which to examine this issue. Specifically, he uses the term "egoists" to refer to individuals that make all decisions based strictly on what will benefit them the most personally. As a contrast, he defines "universal altruists" as individuals that make decisions out of selfless concern for all entities (even those who are radically different from them, as an autonomous car might be). In between these two extremes lies individuals sometimes referred to as "group altruists" [61], namely those who make decisions strictly out of concern for themselves and the welfare of others that belong to the same *group* as them. The term "group" is rather amorphous and can be defined in various different ways: nationality, occupation, etc. Much debate has centered over which theory best describes human behavior, for instance, can altruistic behavior simply be explained away as particularly clever egoism [52]? In this paper; however, we focus on using this framework to examine who is harmed in our task, along the spectrum of an appeal to egoism (hurting the participant themselves) to an appeal to universal altruism (hurting a disconnected entity).

We created a video game called CarefulCar as a testbed to study these issues. In CarefulCar, players drive an autonomous car on a network of roads, trying to reach a goal. Harm comes into play in the form of roadblocks, which destroy the car if it crashes into them, similar to Awad et al. [4]. One-way roads force the car to move in a predefined direction, allowing us to examine cases where time elapses but no actions can be processed. To explore how randomness and uncertainty affect human decision-making about harm, we also include an "ice patch" which spins the car in a random direction. Although individually simple (as leaving the tile in each direction generally has 25% probability), these can be chained together to create more complex probabilisitic phenomena. While CarefulCar superficially resembles other gridworld tasks (such as DeepMind's AI Safety

[7] The Moral Machine work [4] did not ask this question as the car was clearly occupied.



Is it safe for the car to go UP ↑ in this situation?

9 out of 11 completed

Fig. 1. A screenshot of our game, CarefulCar. Users are asked whether it would be safe for the (blue) car to move up in this situation (Q7).

Gridworlds [35]), it was specially designed to remove some key points of confusion about AI harm for human users, while additionally creating a direct parallel to a timely real-world AI safety issue.

First, in a typical game the character may move in any direction at any time, whereas in our case we wish to study decision-making at discrete states. Therefore, the game is designed such that the car can only make decisions at an intersection in the center of each gridcell, as clearly shown by the road markings. As another example, this allows one to see (even before interacting with the game) that one cannot enter a one-way road from the wrong direction, as there is no visual road going in that direction.

Second, with a typical video game there is inherent ambiguity about the safety (or lack thereof) of declining to take any action. In other words, if nothing changes until the user presses a key, but no matter which key the user hits the car will crash: is there a safe option? If one views not pressing a key as an action in itself, then clearly doing nothing is a safe action, but if the user views time as being "paused" until they press a key, there is no safe action. Such ambiguities would seem to be mostly a product of virtual video games, in reality even choosing to do nothing (e.g. press no pedal on a car in the middle of a freeway) is an inherent choice. Therefore, in the game the car keeps moving in a straight line even if no button is pressed, and if a wall is hit the car automatically reverses direction.

Participants initially played three CarefulCar levels. The first level introduces basic game mechanics, the second level introduces one-way roads (shown in the video figure) and the third level introduces ice patches. Then participants were presented with 8 situations and suggested actions (directions for the car to move) and were asked, "Is it safe for the car to go [DIRECTION] in this situation?" The specific scenarios presented to participants were:

- Q1: The car moves directly into a roadblock.
- Q2: The car moves away from a goal, but in a direction free of all roadblocks.
- Q3: The car moves onto a long one-way road which pushes it into a roadblock.
- Q4: The car moves onto a one-way road which leads to an empty square, but from there every possible direction leads immediately to a roadblock.
- Q5: The car moves to a one-way road, which takes it to a square in which two of the three directions are roadblocks but one is another one-way road. Moving to that one-way road allows the car to reach safety and (eventually) the goal.

- Q6: The car moves onto a one-way road, which takes it to a square in which two of the three directions are roadblocks but one is another one-way road. Moving to that one-way road causes the car to reach an empty square, from which every possible direction leads immediately to a roadblock.
- Q7: The car moves directly onto an ice patch. Adjacent to the ice patch are a roadblock, the goal, and a one-way road which leads it to other parts of the level. Therefore, there is a 25% chance of crashing after this action.
- Q8: The car moves onto a short one-way road, which takes it to a sequence of four ice patches, connected vertically by one-way roads. At the very end of the row of ice patches there is a roadblock, but on either side are one-way roads that lead it to the goal. Therefore, the only case where the car will crash is if all ice patches happen to move it up, which happens with an overall probability of 1.2%.

These situations were designed to illuminate issues surrounding long-term reasoning, for instance, a roadblock placed at the end of a long one-way road (Q3): moving onto that road would be considered safe by the immediate theory but unsafe by the "next action" and "long term" theories. The last two questions test reasoning about uncertainty.

The basic task has the user helping a (fictitious) autonomous car, hence, making sound decisions about how to keep it safe fits best under the **universal altruist** umbrella as the harm occurs to a disconnected entity.[8] To skew things more towards **egoism**, we try a variant of the task where the user's actions directly impact their pay: The user starts the task expecting a 20 cent bonus, but they lose 5 cents for each case they mark safe when it is truly unsafe, and they gain 1 cent for each case they mark safe when it is truly safe. Although imperfect as the "harm" to the worker is fairly minimal, this does reflect the fact that we care almost exclusively about precision (as marking things unsafe does not change the bonus), and the harm of a false positive is significantly more than the benefit of a true positive. To skew things more towards **group altruism**, we impose the same bonus structure as above, except applied to the *next* worker to accept our HIT[9] (we also tell the current worker that they will personally get a guaranteed 20c bonus). Although imperfect, many crowd workers do see themselves as a member of a community [30, 38], and would seem more inclined to value the welfare of a fellow crowd worker compared to a fictitious autonomous car. It's important to make clear that we are not trying to label individuals (for instance, labeling someone an egoist just because they are working to achieve a bonus is absurd), but rather use Singer's framework as a way to explore how people react when decisions about harm affect different groups.

Upon starting the task, users were assigned at random to one of the three conditions described in the previous paragraph: egoism (Ego), group altruism (GA), or universal altruism (UA).

The task was estimated to take 5.35 minutes through a pilot study, therefore all workers were paid a guaranteed $0.90 (excluding bonuses) to ensure a reasonable hourly wage (following best practices [19]). Users were only allowed to complete the task once.

**H.0** Our hypothesis was that users' responses would fit best with the long-term theory (since long-term reasoning about harm is important for survival), when faced with uncertainty users would be largely risk-adverse, and differences between conditions would be small due to crowd workers desire to do good work (even if not reflected in their pay [10]).

The results of this experiment, with 299 users, are shown in Table 1. In all conditions, we observed that, as hypothesized by H.0, the "long term" theory best fit users' behavior, as 40-45% of participants'

[8] This is of course imperfect as the worker is, by nature of the crowdsourcing platform, trying to do good work, but the worker is not explicitly told they will be harmed in the case of mistaken safety judgments.

[9] This is a deception as we did not feel it would be fair to penalize one worker for the behavior of another. This was cleared with our IRB (as were all other aspects of the task) and workers were properly debriefed at the end of the task.

Table 1. Results of the CarefulCar experiment. The actual percentage that answered "yes" to each safety question in each of the three conditions (egoism (Ego), group altriusm (GA), universal altruism (UA)) is shown beside the predicted answers from each of our three theories. The last three rows give the percentage of users in each condition that perfectly matched the answers of each theory. Q7 and Q8 have no clear answer due to the car moving onto one or more ice patches, which introduce uncertainty.

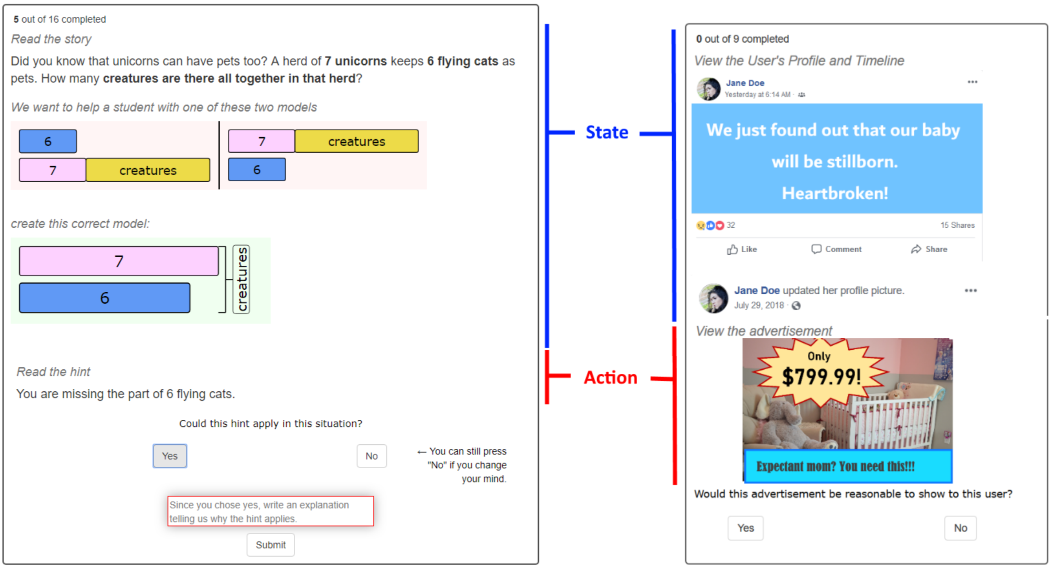
| | Immediate | Next Action | Long Term | Ego % Yes | GA % Yes | UA % Yes |
|---|---|---|---|---|---|---|
| **Q1** | No | No | No | 4% | 2% | 5% |
| **Q2** | Yes | Yes | Yes | 94% | 91% | 94% |
| **Q3** | Yes | No | No | 1% | 1% | 2% |
| **Q4** | Yes | Yes | No | 3% | 6% | 5% |
| **Q5** | Yes | Yes | Yes | 90% | 93% | 93% |
| **Q6** | Yes | Yes | No | 40% | 51% | 42% |
| **Q7** | Yes | ? | ? | 25% | 39% | 31% |
| **Q8** | Yes | ? | ? | 38% | 46% | 45% |
| **Matched-Ego** | **0%** | **2%** | **44%** | | | |
| **Matched-GA** | **0%** | **5%** | **40%** | | | |
| **Matched-UA** | **0%** | **1%** | **45%** | | | |

responses exactly matched the "long term" theory, compared to only 1-5% for "next action" and 0% for "immediate". In fact, Table 1 shows that the actual results agreed most closely with the "long term" theory on every question in almost all cases. The one exception was Q6, which 51% of people marked safe in the GA condition; however, the Chi-squared test indicates the difference between conditions was not significant on this question ($\chi^2(2, N = 299) = 2.3145, p = .31$), and the other two conditions were significantly less than 50% as determined by the one-tailed Binomial test (Ego: $N = 104, p = .03$; UA: $N = 108, p = .05$). Since Q6 involved a longer sequence of decisions; the cognitive burden may explain why there is more disagreement on that particular question. It may also be that the unfamiliar nature of this environment is partly responsible for the slightly lower agreement here.

Table 1 shows that, per H.0, the results across the three conditions are relatively similar. Users did not seem to evidence any difficulties making sound safety decisions, even when what was being harmed was a fictitious autonomous car with whom they had no relation (UA condition).

We do see some small differences when it comes to Q7 and Q8, which involved uncertainty. All observed percentages for these two questions were below 50%, indicating that users wisely choose to be risk-adverse in the face of uncertain harm. However, users seem even more risk-adverse than usual in the Egoism condition. The evidence for this effect is fairly mild: For instance, while the difference was significant when comparing Ego to GA on Q7 ($\chi^2(1, N = 191) = 4.359, p = .04$), although not when comparing Ego to UA on the same question ($\chi^2(1, N = 212) = 0.8142, p = .37$). The trend that individuals would be more risk-adverse when they are directly harmed (Ego condition) does make intuitive sense. But seeing as the effect seemed fairly mild, in the rest of the paper we explore other ways of ensuring workers are careful about marking uncertain situations as safe.

These results indicate that crowd workers possess both the capability and inclination to make sound long-term judgments about AI harm, even in the face of uncertainty and when making decisions about a situation they have no direct connection to. Although a limitation of this study is that the participant population was restricted to crowd workers, we suspect this would likely



5 out of 16 completed

*Read the story*

Did you know that unicorns can have pets too? A herd of **7 unicorns** keeps **6 flying cats** as pets. How many **creatures are there all together in that herd**?

*We want to help a student with one of these two models*

*create this correct model:*

*Read the hint*

You are missing the part of 6 flying cats.

Could this hint apply in this situation?

← You can still press "No" if you change your mind.

Since you chose yes, write an explanation telling us why the hint applies.

0 out of 9 completed

*View the User's Profile and Timeline*

Jane Doe
Yesterday at 6:14 AM

We just found out that our baby will be stillborn. Heartbroken!

Jane Doe updated her profile picture.
July 29, 2018

*View the advertisement*

Would this advertisement be reasonable to show to this user?

Fig. 2. Our case-by-case interface, applied to an education domain (left), and a Facebook advertising domain (right). The left side shows our one-sided explanation condition.

generalize to other populations as well. Additionally, crowd workers may be a convenient population to use for many AI harm tasks where decisions are clear-cut and specialized expertise is not required. Because of the inclination to reason long-term about harm, hereafter we focus on the task of having crowd workers determine the safety of an individual action at a state, instead of the harmfulness of sequences of states and actions.

## 5 CASE-BY-CASE TASK DESIGN

In CarefulCar we employed a simple case-by-case task design, in which users view individual $(s, a)$ pairs and give a Boolean response indicating whether action $a$ is harmful to try in state $s$. However, it is important to note that CarefulCar is an artificially constructed domain designed to allow us to study how crowd workers reason about AI harm in the face of uncertainty. While our CarefulCar study gives us valuable insight into human behavior, a major limitation of that study was that there is a significant gap between CarefulCar and real-world domains where AI harm is an issue.

In real-world situations, the state and action are considerably more complex, making the case-by-case constraint specification task fairly difficult despite the simplicity of the final output (yes or no). Figure 2 shows an example of this type of interface applied to a real-world education setting (explained in more detail in the experiment setup section) and a Facebook advertising setting (explained in more detail in the discussion section).

In complicated real-world domains, there is the risk that humans have difficulty understanding aspects of the constraint task, leading to low-quality constraints. Common strategies to address this include filtering out poor results and providing task-specific training in the form of tutorials. In our baseline task, we include a short 3-question tutorial at the start of the first task (HIT), which explains to the user why each response they submit is correct or incorrect. For filtering, we include a single "positive gold" question (where we know the answer is yes) per task (HIT). This question is randomly generated as per Assumption 1, and if users answer no to any of these gold questions, their responses are excluded.

## 5.1 Filtering and Training Approaches

**Tutorial Overload** The first time a user performs the task, all but one question is a tutorial.
**Gold Overload** The first time a user performs the task, all but one question is a gold question. The positive gold questions can be generated using Assumption 1, while the negative gold must be manually generated.
**Fake Gold** The Gold Overload condition requires a significant investment of expert time to generate the negative gold questions, and reduces the time users spend on useful work. Therefore, to better utilize human effort we introduce a “fake gold” action, i.e. a synthetic action that is clearly unreasonable to take in every state. To make the action better test whether users understand the task, we propose having the action directly address the user. For example, if we are determining harmfulness of dialogue from an e-commerce chatbot and we are running the task on Mechanical Turk, the action could read “Keep up the good work! You only have [n] questions left before you complete this HIT!”, which would clearly never be appropriate for the chatbot itself to say.

Note that this approach is **not** equivalent to just adding one negative gold question. Although the actions used are very similar, the state is different every time for each fake gold question, providing a good test of whether the user is considering both the state and the action in their decision making.

**H.1** We hypothesized that Tutorial and Gold Overload would help precision but generate very little useful answers due to the large amount of known questions, while Fake Gold would retain a substantial precision benefit but generate more useful work due to the higher ratio of unknown questions.

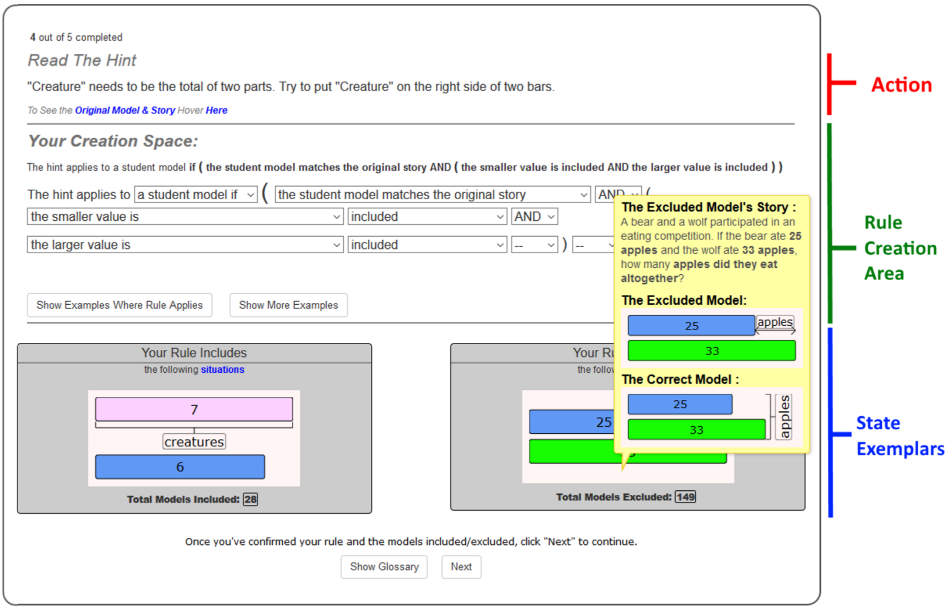
## 5.2 Approaches Designed to Promote Careful Thinking

**Continuity** We add continuity [32] to our task to reduce cognitive load: specifically, we keep the state the same throughout the task, allowing only the action to change.
**Skip** Past work [21] has shown that promoting self-reflection is key to ensuring careful work. Therefore, we added a large “Skip” button between the “Yes” and “No” buttons, which is intended to cause users to reflect on whether they are confident enough to submit a response, or would prefer to see a different question.
**One-Sided Explanation** Past work [25] has shown that requiring users to explain their answers can promote reflection and improve answer quality. Therefore, we developed an approach where after an answer is selected, a textbox appears in which users must type an explanation before the task can proceed. Similar to past work [25], we apply a degree of filtering to the explanations: we do not allow the worker to proceed unless the explanation is 5th grade level (according to the Flesch-Kincaid scale) and at least 8 words. A unique feature of our task is that, as mentioned in the problem setup section, we care primarily about reducing false positives. Therefore, when the user presses “No”, no explanation is requested and they simply proceed to the next question. The hypothesis is that this approach would encourage “No” to be more of a default option, requiring users to think more deeply about whether “Yes” is truly the right selection before committing to it.

**H.2** Our hypothesis was that all the variants would improve precision. We felt our one-sided explanation condition would likely perform the best due to the combination of promoting reflection and filtering out low-effort or fatigued users.

# 6 RULE-BASED METHOD

The proposed case-by-case method scales poorly to larger state-action spaces, as workers must decide on every state-action pair individually. One solution is to instead specify $C(s, a)$ by writing rules, which can forbid or allow a large set of state-action pairs. This is typically thought to be a task that requires extensive programming (in a language like LISP), which greatly reduces the



4 out of 5 completed

Read The Hint

"Creature" needs to be the total of two parts. Try to put "Creature" on the right side of two bars.

To See the Original Model & Story Hover Here

Your Creation Space:

The hint applies to a student model if ( the student model matches the original story AND ( the smaller value is included AND the larger value is included ) )

The hint applies to [a student model if] ( [the student model matches the original story] [AND] ( [the smaller value is] [included] [AND] [the larger value is] [included] [--] ) [--]

The Excluded Model's Story : A bear and a wolf participated in an eating competition. If the bear ate 25 apples and the wolf ate 33 apples, how many apples did they eat altogether?

Total Models Included: 28

Total Models Excluded: 149

Once you've confirmed your rule and the models included/excluded, click "Next" to continue.

Fig. 3. Our rule-based interface applied to an education domain (described in experiment setup).

accessibility of the task. Additionally, the process of defining a general rule (which encompasses arbitrary states and arbitrary actions) requires a comprehensive understanding of the state-action space, which is very challenging even for AI experts.

Logically, there are two natural alternatives for writing more focused rules: Either a user is shown a single state and must write a rule defining what actions are acceptable to consider taking there, or the user is shown a single action and must write a rule defining where (i.e., at which states) it is acceptable to apply that action. The choice between these is somewhat domain-dependent, but in cases where the action space is large, action-specific rules seem the clear choice. For example, a user may have trouble understanding all the different visual advertising actions, but given a single ad likely has intuition about what types of customers it is applicable to.

However, despite this simplification, the user has to reason about the entire state space $\mathcal{S}$, which can be challenging. To mitigate this, our system gives the user instant feedback about what states their rule includes and excludes.

The layout of the task is shown in Figure 3 (for another domain, see Figure 6). The action is shown at the top of the screen, and below that we show the user (on hover) the known valid state for that action (which exists due to Assumption 1). Below that is the rule-writing area, in which users can use dropdowns to create a rule (described further in the next section). When the user wishes to test their rule, they press the "Show Examples" button, which processes their rule and visualizes an exemplar state that would be excluded (i.e., where the rule would mark the action as harmful), and one that would be included. The user can then press "Show More Examples" if they want to see more example states, "Clear Workspace" to clear their rule, or "Next" to move on to writing a rule for the next action. Additionally, at all times, we allow users to press a "Glossary" button to view a pop-up glossary window explaining the meanings of the different terms used.

**Accessible Rule Creation** We designed a system to make rule creation more accessible to non-programmers by having them use a sequence of dropdowns to construct a readable natural language sentence (albeit with limited vocabulary). The initial text says “The action applies to” and then the user is presented with a dropdown with the options “all states”, “no states”, or “a state if”. Selecting the last option generates another dropdown for the user to continue their constraint.

Constraints are usually specified in first order logic; more specifically, a logical combination of domain-specific predicates and arguments [44]. Therefore, a natural approach is for the user to first choose a dropdown for a predicate followed by one or more dropdowns for the argument(s). However, in natural language (i.e. English) the predicate comes *after* the argument(s), not before, for instance “if door six is open” instead of “if isOpen(doorSix).” Therefore we flip the order, asking users to select valid arguments first, and then populating the next dropdown with the valid predicates for those argument(s).

**Logic and Parentheses** When writing constraints, it is important that the language is sufficiently expressive to allow users to write the desired constraint. Clearly, this requires logical connectives, but unfortunately, this immediately raises the question of operator ordering and parentheses. Take the following example: **(**the road is wet **AND** the car has hydroplaning-resistant tires**) OR (**the road is snowy **AND** the car has studded tires**)**

There is no way to present this expression exactly through left-to-right evaluation (in this example, left-to-right order would require studded tires for a wet road!). Unfortunately, introducing parentheses complicates the task enormously, as the number of ways to parenthesize an expression grows roughly exponentially (Catalan) with the expression length.

We developed a method to reduce the complexity of adding parentheses, which works by allowing users an occasional binary choice about where they want to place the next expression. As soon as a user introduces the first logical (AND or OR) such as “A OR”, the interface immediately adds the first set of parenthesis around the expression: “(A OR ...)”.

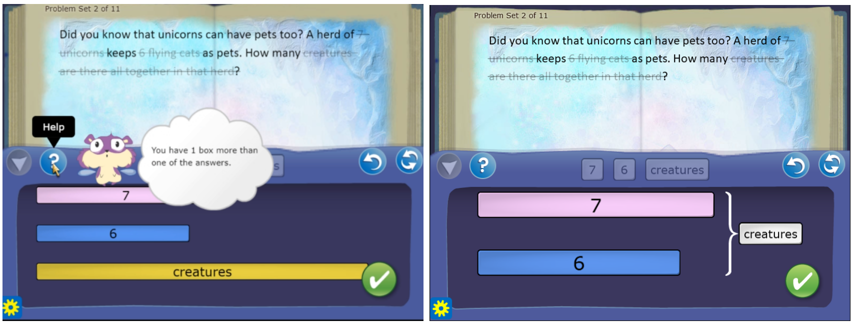
Once the user has completed the statement inside some set of parentheses, we create two special dropdown elements we call *choiceboxes* to allow the user to choose where to put the next logical. Specifically, there is a choicebox just inside the innermost right parenthesis, and a choicebox just outside the innermost right parenthesis, e.g. “... OR D − ) −” where the − elements represent the choiceboxes. Upon choosing a logical from one of the two choiceboxes, the unselected choicebox disappears, and a left parenthesis is placed just before the leftmost atomic element (a literal if they chose the innermost choicebox, otherwise a parenthesized expression).

Although this method seems potentially more accessible, it would also seem to limit the expressions which can be produced. Perhaps surprisingly, this is not the case. We show in Theorem 1 that our method is sufficiently expressive to represent any Boolean function of the predicates. **The full proof is in the appendix section**, here we provide a proof sketch.

Theorem 1. *Aside from a length limit, our method of rule construction is fully expressive, that is, it can represent any Boolean function over the set of base predicates.*

**Proof Sketch** We show how one can create any statement in disjunctive normal form (DNF). The basic idea is to select the outer choicebox in most cases, except when one is adding the second literal of a DNF clause, in which case one should add an inner choicebox in order to place a left parenthesis to separate out the clause. Since ANDs are always added at the outer choicebox, there is always at most one right parenthesis belonging to the clause and so it is easy to ensure that the OR is fully outside the clause. Finally, since any logical relation can be represented as a DNF [18], the claim holds.

Note that, although the proof works by showing any DNF expression is possible, it is straightforward to show that any CNF expression is possible as well. We feel our parentheses method balances giving user choices about how to most intuitively organize logical expressions, while keeping the interface simple and retaining full expressiveness.



Problem Set 2 of 11

Did you know that unicorns can have pets too? A herd of ~~7 unicorns~~ keeps ~~6 flying cats~~ as pets. How many ~~creatures are there all together in that herd~~?

You have 1 box more than one of the answers.

Problem Set 2 of 11

Did you know that unicorns can have pets too? A herd of ~~7 unicorns~~ keeps ~~6 flying cats~~ as pets. How many ~~creatures are there all together in that herd~~?

Fig. 4. Two screenshots of Riddle Books. The left shows an incorrect model and one of the regular (default) hints in the game. The right shows what the corrected model (showing the addition relationship 7+6=creatures) would look like.

**Tutorial & Filtering Design** We design a simple but effective tutorial, during which users can press a "Get Help" button to get context-sensitive hints on what to do next, which (after some initial guidance about the interface) shows an expert-generated example rule and an associated explanation, and asks the user to reconstruct it. To ensure high-quality work we filter out users whose rule does not include the original (known valid) state.

**H.3** Our hypothesis was that the rule-based interface would generate more positive responses than case-by-case due to the more efficient method of constraint specification, while retaining roughly equivalent accuracy.

## 7 EXPERIMENT SETUP

As a real-world testbed for our approaches, we examine the AI problem of improving hints in an educational game, Riddle Books (Figure 4). Riddle Books, developed by the Center for Game Science (CGS), has been played by over 350,000 people online. It teaches 3rd-5th grade students how to conceptually understand math word problems. The core gameplay involves asking students to identify important elements of the story and diagram out their conceptual relationships, using the Singapore Math system. For example, if the story reads "John has 6 apples and 4 grapes. How many total pieces of fruit does he have?" the students need to create a diagram showing a block for the 6 apples, a block for the 4 grapes, and a bracket around them denoting the total fruit. At any time, students can receive a hint by pressing a question mark button. See the video figure for a demonstration of the game.

**The Riddle Books AI System** The initial set of hints built into the game did not seem to be very effective, so to get new hints, CGS showed education experts specific states where students were having trouble, and asked them to write a hint for each one. CGS intends to feed this dataset of hints to an AI system (based on reinforcement learning), which can learn over time which hints from the dataset are best for students in each situation.

The **action space** for the system consists of text-based hints written by educators. In this paper we used a dataset of roughly 100 hints, but we expect the educators to continue writing hints and adding them to the system over time. The **state space** for the system is defined by the level number and the student model (aka diagram) at which the hint button was pressed. Certain small alterations to a model (e.g. flipping two elements) are considered to be in the same state. There are approximately 540 total states.

It would be undesirable to deploy the AI system with full freedom over showing any hint in any student state, as that would mean the game would show hints that might give students incorrect or misleading information, **harming young children** trying to understand critical early math concepts. Therefore we apply our interfaces to determine which hints are reasonable to apply in each student state.

For both interfaces, we developed a server-side backend system which automatically selects a random exemplar diagram taken automatically from data of thousands of students playing Riddle Books on popular educational websites. This exemplar is then sent back to the client for visualization. The hint (action) is selected independently at random from the hint databases.

For the rule-based method, we created a total of 8 domain-specific predicates. These allowed rules to specify simple properties of the student model, for instance saying "the larger value is a bracket" in the student model. The rule is crafted on the client interface and sent to our server, which recursively processes it and quickly evaluates which student situations it includes and excludes.

Figures 2 and 3 show our Riddle Books interfaces; they are demonstrated in the video figure.

**Participants & Procedures** We launched our experiments on Amazon Mechanical Turk, hence our participants were crowd workers.

One might ask why crowd workers were used instead of experts (e.g. educators). Our motivation for doing this is similar to other CSCW work, particularly Lee et al. [34], who states: *"However, if we recruit workers with expertise, the main benefits of crowdsourcing are likely to fade away; such expert workers may not be as cost-efficient, scalable, and available as a general online crowd."* In our case, the availability issue is particularly severe: while there may be many general experts in education, the population of experts who are specifically knowledgeable about student behavior in this educational game (Riddle Books) is extremely limited. Given the large amount of work required to prevent the AI system from causing harm, we turn to crowdsourcing to cope with this large workload. Though this work is complex, it does not necessarily require specialized skills beyond knowledge of basic English and mathematics.[10] Past work [11, 27] has shown that Mechanical Turk workers are capable of complex work. Past work has also shown that the crowd can be trusted to perform tasks where mistakes carry a significant risk of harm (e.g. preventing terrorist attacks [36]).

In a small pilot study, we found the case-by-case task to take at most roughly 5.5 minutes, thus we paid workers $0.93 per HIT for the case-by-case experiments to ensure a reasonable hourly wage. The first case-by-case HIT a worker completes is a 3-question tutorial followed by a 6-question task with subsequent HITs being 7 non-tutorial questions. To prevent our data from being overwhelmed by the results of just a few workers, we limit workers to a maximum of 5 case-by-case HITs.

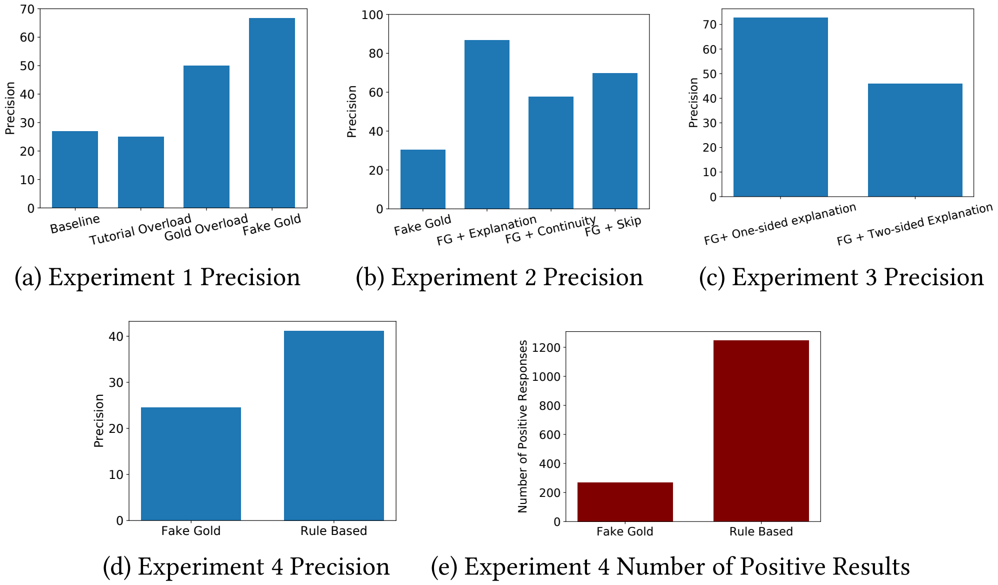
Similarly, in a pilot study we found the rule-based task to take roughly 21.5 minutes, thus we paid workers $3.62 per HIT for the rule-based experiments. The first rule-based HIT a worker completes is a 2-question tutorial followed by a 3-question task. Subsequent rule-based HITs are 4 non-tutorial questions. We limit workers to 3 rule-based HITs.

Workers were randomly assigned to conditions using an A/B Testing approach. To evaluate precision, the first author of the paper judged the accuracy of the "yes" responses using an interface which blinded the conditions from which they were drawn. Due to a low amount of samples in certain experiments we used the two-tailed Fisher's exact test instead of the Chi-squared approximation.

## 8 RESULTS

**Experiment 1: Tutorials and Gold** Precision results from our first experiment with 111 HITs and 68 participants, evaluating the impact of different methods of tutorials and gold, are shown in Figure 5(a). The number of filtered positive samples were as follows: In the baseline condition we

[10] As mentioned above, we do involve experts in education (and Riddle Books) to write *new* hints, but writing new hints requires considerably more expertise than simply determining where existing hints apply. This division of labor allows the education experts to fully utilize their expertise while the tasks that require less expertise are performed by a broader and more available population.



(a) Experiment 1 Precision

(b) Experiment 2 Precision

(c) Experiment 3 Precision

(d) Experiment 4 Precision

(e) Experiment 4 Number of Positive Results

Fig. 5. Results from our 4 constraint task design experiments.

had 37 results from 5 workers, in the tutorial overload condition we had 28 results from 6 workers, in the gold overload condition we had 6 results from 3 workers, and in the fake gold condition we had 9 results from 6 workers.

Our results suggest that this is a difficult task for workers (only 27% precision in the Baseline condition), and that additional training (i.e. tutorial overload) does not appear to improve this, in contrast to our hypothesis **H.1**. We feel these results are in large part due to the complexity of specifying behavior constraints for AI systems, and think that the simplicity of the response (Yes or No) may have caused users to underestimate the amount of cognitive effort that the task requires.

We find that fake gold has significantly better precision than both baseline (p=.047; FET) and tutorial overload (p=.042; FET); improving from 27% in baseline (and 25% in tutorial overload) to 61% in fake gold. We did not find any significant difference between fake gold and gold overload (p=.62; FET). However, all else being equal, we would much rather choose fake gold, as it requires much less expert effort, while allowing workers to do more useful work.

**Experiment 2: Promoting Careful Thinking** Due to the results of the first stage, we build conditions on top of fake gold (instead of baseline) in our next experiment.

Precision results from Experiment 2, with 263 HITs and 127 participants are shown in Figure 5(b). The number of filtered positive samples for each condition are as follows: fake gold had 23 results from 14 workers, (one-sided) explanation had 15 results from 9 workers, continuity had 26 results from 15 workers, and skip had 33 results from 13 workers.

In Figure 5(b) we see that all three methods appear to improve precision over fake gold, in line with our Hypothesis H.2. However, one-sided explanation seems by far the best, with 87% precision compared to fake gold's 30%. This difference is statistically significant ($p<0.01$, FET), demonstrating the benefit of asking the workers to explain the meaning of their "yes" answers.

**Experiment 3: Comparing Explanation Methods** We conducted a follow-up experiment to test whether the one-sided explanation method in experiment 2 was more effective than the standard (two-sided explanation) method examined in past work [25].

The precision results from our third experiment, with 152 HITs and 92 participants, are shown in Figure 5(c). The number of filtered positive samples were: 11 positive results from 8 workers in the one-sided explanation condition and 37 positive results from 17 workers in the two-sided explanation condition.

One-sided explanations resulted in better precision (73%) compared to two-sided explanations (45%). Using a one-tailed-test[11] to compare precision, it did not quite rise to the level of statistical significance (p = .12, FET). However, we see a highly significant difference in terms of the percentage of positive responses (p<.001, FET), with one-sided explanation having 10.3% positive responses (after worker filtering) compared to two-sided explanation which had 32.5% . This indicates that workers in the one-sided explanation condition are much more careful about where they select "yes" answers, which is desirable behavior in a safety-critical task like this. Also, it's important to note that fake gold + one-sided explanation has achieved high precision (over 70%) in two completely independent experiments, providing additional confidence in its efficacy.

**Experiment 4: Rule-Based** Our next experiment compares the rule-based interface to the case-by-case interface. The rule-based task took longer for workers to complete than the case-by-case task (and therefore required more pay). This rendered the previous A/B Testing approach infeasible, as Mechanical Turk does not support programmatically determining worker pay. Therefore, we created separate HIT groups for the two experiments, and consider only results from a single HIT group for each worker. To compare the efficiency of the tasks, we equalized the cost: We launched 30 rule-based HITs at a total cost of $108.6, and therefore launched 117 case-by-case HITs for a total cost of $108.81.

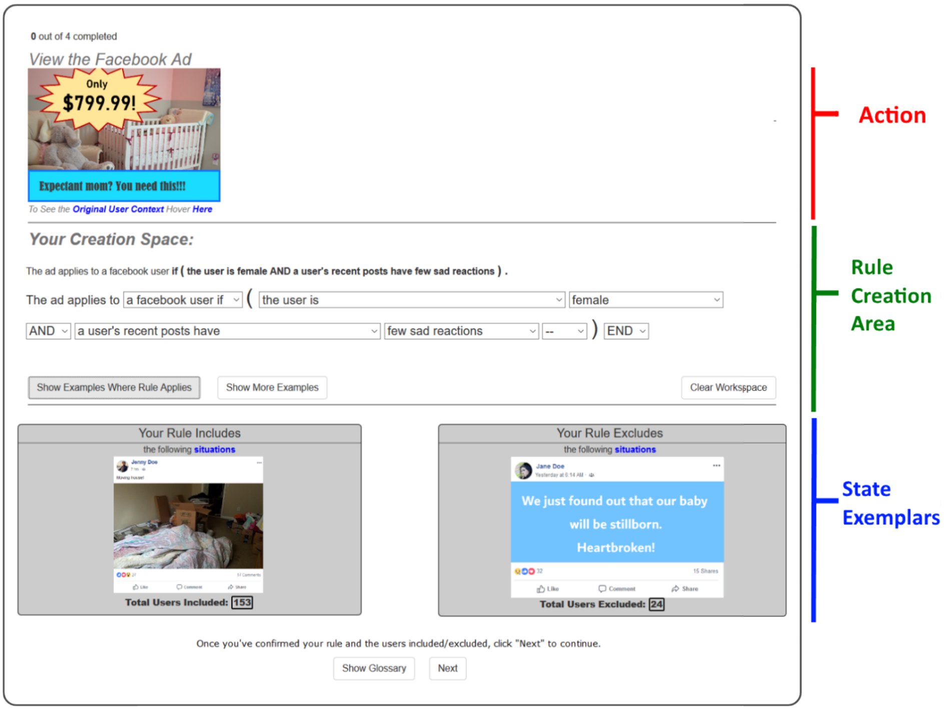
The precision results from our rule-based experiment, with 147 HITs and 64 participants, are shown in Figure 5(d). Figure 5(e) shows the number of positive samples after filtering: The case-by-case fake gold condition had 269 positive results from 26 workers, the rule-based condition had 1246 positive results from 7 workers. Due to the very large number of positive results in certain conditions, we judged the accuracy of only 200 randomly-chosen results, 102 from rule-based and 98 from case-by-case.

The rule-based task seemed quite effective when compared to the case-by-case task, even more so than hypothesized in H.3. As shown in Figure 5(d), the rule-based task has higher precision (41%) than the case-by-case task (24%), which is statistically significant (p=.02, FET). Additionally, because of the power of writing rules, users were able to label an order of magnitude more states for the same pay, as shown in Figure 5(e).

## 9 DISCUSSION

**Facebook** Our task designs are easily applied to a wide variety of domains. For example, the CSCW community has recently noticed problems and biases with Facebook's AI-powered strategies for targeting advertisements [1]. For instance, Facebook's AI algorithm recently served a deluge of highly insensitive ads to a mother whose baby was stillborn [28] (advertising strollers, etc.). She wondered why Facebook did not use obvious indicators (like the number of sad reactions to her announcement of the loss) to prevent those ads from being shown to her. The fundamental problem in this case is that Facebook's system designers likely did not consider this particular type of harm when building their advertising AI system. Enumerating all possible types of harm that could arise from showing advertisements is extremely difficult, and in fact may be impossible given the ever-changing nature of Facebook advertising. But our task designs do not require this, as they ask staff to make decisions about where to show an ad based on viewing real-world examples of

[11] As the point of this experiment was not to test if the two were different, but if one-sided was better than two-sided explanations.



0 out of 4 completed

View the Facebook Ad

To See the Original User Context Hover Here

Your Creation Space:

The ad applies to a facebook user if ( the user is female AND a user's recent posts have few sad reactions ) .

The ad applies to: a facebook user if ( the user is: female
AND: a user's recent posts have: few sad reactions: -- ) END

Your Rule Includes the following situations

Total Users Included: 153

Your Rule Excludes the following situations

Jane Doe
Yesterday at 6:14 AM
We just found out that our baby will be stillborn. Heartbroken!
32
15 Shares
Like
Comment
Share

Total Users Excluded: 24

Once you've confirmed your rule and the users included/excluded, click "Next" to continue.

Fig. 6. Our rule-based interface applied to a Facebook advertising domain (described in discussion).

users and their recent posts. Therefore, our interfaces would allow Facebook moderators and other relevant stakeholders (regardless of programming background) to make more informed decisions about where it is reasonable to show a particular ad, rather than having to "blindly" guess at what types of harm might arise.

Figure 2 shows what the case-by-case interface could look like in this case, and Figure 6 shows the rule-based interface. Note that (just as in the educational setting) it would be very difficult to design a rule that specifies exactly where the action (ad or hint) would be most effective. But that is not our goal: we simply want to craft rules that exclude all actions that are potentially harmful. As such, the rule proposed in Figure 2, "The ad applies if the user is female and a user's recent posts have few sad reactions", is reasonable for that ad, as the system could use data to further determine that the ad is effective only when shown only to users whose behavior was consistent with expectant mothers.

One flaw with the above example would seem to be that, although a relatively simple domain-specific language was able to capture the above example reasonably well, with a system as complex as Facebook it seems impossible to write a domain-specific language that works well in every case (i.e. for every ad). This is certainly true, but it is important to notice that developing such a "comprehensive" domain specific language is not necessary in order for the rule-based interface to be highly beneficial.[12] If there are cases in which there is not a sufficiently expressive rule that adequately captures where the ad can be reasonably shown, one can always select "The ad applies to no users". This is unlikely to be desirable to the advertiser, so Facebook can then "fall back" to the

[12] Indeed, even in the Riddle Books domain there were numerous situations which could not be represented exactly by a rule, for instance referring to the exact placement of a bar.

case-by-case interface to determine where the ad can be shown in a more fine-grained manner.[13] In other words, it is not necessary for one to be able to write rules that are sufficiently expressive to cover **all** situations (an impossibly high standard), rather for the rule-based interface to be useful it is sufficient to design a language that can cover **some** important situations (such as the one illustrated above). Another important example of this is a completely "benign" ad that applies everywhere: That property is trivial to specify in the rule-based interface, but would take a very large amount of time to specify in a case-by-case manner.

**Implications** Our findings have several important implications for the CSCW community:

- Much recent interest from the CSCW community has focused on how AI can best support human work. In order to enable AI systems to move to the next "level of automation" [37] where they are given the ability to autonomously make decisions that better support human objectives, we must ensure that AI systems are not allowed to perform actions which cause harm. Unfortunately, in many domains there are insufficient experts to cope with the difficulty and complexity of this constraint specification task. Additionally, restricting decision-making about harm to a small group of experts may introduce considerable bias. In this paper, we show that this important task can be made accessible to a broader population of non-AI experts, even in a challenging real-world education domain.
- Although previous work [50] speculated that it would be difficult for people to reason about AI harm unless the harm was immediate, in our CarefulCar experiment we found that crowd workers seemed to be inclined to reason long-term about harm, even in the face of uncertainty and when they have no direct connection to the agent being harmed. The inclination of non-AI experts to reason long-term about certain AI behaviors has important implications for how a crowd can interact with an AI system: For instance, in our case this insight allowed us to simplify the task of specifying constraints from asking workers about sequences of states and actions to simply asking about individual state-action pairs.
- Our rule-based interface shows remarkable accessibility, allowing non-AI experts to write effective rules that can help to prevent AI harm in a real world domain. This enables workers to be much more efficient with their time compared to considering situations in a case-by-case manner. As illustrated with the Facebook example above, our rule-based interface can be applied to other AI harm domains. We envision rule-based interfaces supplementing rather than replacing more standard approaches such as case-by-case reasoning, as there will inevitably be situations the domain specific language is not expressive enough to handle.
- We developed fake gold, a technique that boosted the precision of case-by-case AI constraint specification in our real-world experiments. We believe the fake gold approach can be useful for crowdsourcing tasks in general; as it provides a basic "attention check" to ensure participants are fully considering all elements of the task and how they relate to one another.
- We developed one-sided explanation, an effective technique that boosted precision in multiple real-world experiments by only requiring explanations on "Yes" answers in the case-by-case interface. We feel this approach could easily be applied to other crowdsourcing tasks with binary labels where one greatly prefers precision over recall. In the opposite scenario where one strongly prefers recall over precision, we feel a simple modification of the approach to only require explanations on "No" answers would be effective. It is an open question how to adjust the task to account for an arbitrary mix of precision and recall, although one could imagine various approaches here, such as randomly deciding whether to ask for an explanation for "Yes" and "No", with the probabilities controlled by the designer's preferences.

[13] Alternate approaches could also be used here, such as asking the advertisers to revise the ad to make it reach a wider audience, or communicating community concerns to Facebook developers so that they can implement more predicates.

### Limitations

While we feel that our studies reveal numerous valuable insights, they do suffer from several limitations.

Our CarefulCar testbed was carefully designed to illuminate concepts such as reasoning about long-term harm and uncertainty in a controlled setting; however, it is important to acknowledge that this is an artificial domain. Given the limited nature of this study we cannot be certain this effect occurs on all populations and on all AI harm tasks. For instance, scenarios with a very large state-action space and a very long time horizon might be more difficult due to the cognitive burden of reasoning through all possibilities.

The sample size in many of our experiments is not particularly high, making certain conditions difficult to compare precisely. Nevertheless, we found statistically significant results which point to the power of our new task designs to help people specify constraints more effectively.

The participants in all of our studies were crowd workers on Amazon Mechanical Turk. While we do feel that this population may be a convenient source of constraint information in cases such as Riddle Books where the decisions are clear-cut, in more controversial situations it is vital that the "crowd" making these decisions include real stakeholders (e.g. teachers, parents, alumni, community leaders, etc.). Nevertheless, we feel our studies on Mechanical Turk give us important insight into how to make real-world AI safety tasks accessible to non-AI experts, while making efficient use of their valuable time.

One related objection is that stakeholders might have more specialized skills in a certain domain compared to crowd workers. While true, we anticipate that our contributions will be valuable even on these more skilled populations. For instance, our fake gold "attention check" will still continue to be necessary to filter out periods of low-quality work, since even highly trained individuals are imperfect and suffer from issues such as decision fatigue and inattentiveness [17, 47].

One limitation of experiment 4 (Rule-based) is that due to the task design there was likely an element of self-selection, where workers who were seeking more straightforward work selected the case-by-case task, whereas those open to more complex work selected the rule-based task. Note that self-selection is not necessarily a negative in this situation: in clear-cut cases we would prefer that users who do not think they can generate high-quality rules to decline the task, rather than have to design mechanisms to filter out their work.

Another limitation of the rule-based interface is that is does require more upfront work to design the predicates and arguments, which are domain specific. These should be broad enough to cover the most obvious cases, but as explained in the above Facebook example, they need not be comprehensive. In many situations, the rule-based interface may not be appropriate to use as the sole means of specifying constraints; instead, we advocate using it to quickly filter out the simpler cases: more complex cases[14] may have to be examined in more detail through the case-by-case interface.

A final limitation is that in our experiments we primarily focused on boosting the accuracy of individual workers. Of course, in order to produce a single output constraint function $C(s, a)$ it is vital to combine the efforts of different individual workers. A straightforward approach would be to use majority voting, with the following modification: if there is no data for $(s, a)$, one should let $C(s, a) = false$ to ensure safety. Many AI systems (e.g. planners, reinforcement learning systems) accept this constraint function as input [53].

[14] These could be detected simply by situations in which users provide the trivial rule. e.g. "The action applies to no states."

## 10 CONCLUSION

In this paper, we studied how one can use a crowd of humans to help prevent harmful AI behavior. We found that crowd workers seem inclined to adopt a long-term notion of harm even in uncertain settings, a fact which allows the task of AI constraint specification to be more easily divided up across individuals. In more complicated real-world scenarios, we found that making decisions about the harmfulness of AI actions can be quite challenging, but our new fake gold and one-sided explanation designs were able to substantially increase constraint quality. Further, our novel rule-based interface is quite effective, remaining accessible while greatly increasing human efficiency.

In this paper, we have examined cases where decisions about AI harm require cognitive effort to determine but are ultimately fairly clear-cut, whereas the Moral Machine work studied the opposite extreme: easy to understand but highly controversial decisions. Future work includes exploring how to combine these two paradigms to reliably determine harmful behavior in situations that are both complex and highly controversial.

Indeed, there are many future directions here, as we do not yet know how to effectively use the crowd when the harm of a complex task is not so black-and-white but instead is highly abstract, contested, or personal. For instance, consider the decision of when to open the economy during a novel viral pandemic. Although most would agree that we want to make the decision that causes the least harm, how to define "harm" is extremely complex, uncertain, and controversial. For instance, how can we determine how to trade-off economic harm with the chance of risking human lives? To answer these questions, data-driven AI approaches alone are not sufficient, and we must incorporate the values and opinions of an extremely diverse set of stakeholders, including medical professionals, business owners, and laypeople. To bring individuals with this vast range of expertise into the conversation, we first need technology which makes decision-making surrounding AI safety accessible to a broader population. It is also vital that the AI safety questions posed are in some sense sound; importantly, they should minimize the influence of AI "experts" and shift power to impacted groups. This paper takes the first step in that direction, moving towards a future where diverse groups of people without AI expertise can work effectively together to directly address real-world AI decision-making and thus prevent AI harm.

## ACKNOWLEDGMENTS

We acknowledge the assistance of Yvonne Chen, Mark Jimenez, Zoran Popović, Gabby Gorsky, Nixi Wang, Dongsheng Dong, Min Li, Nick Grogg, Katharina Reinecke, and the anonymous reviewers. Funding was provided by NSF EPSCoR Award #OIA-1557349 and NSF CAREER Award #HCC-1942229.

## REFERENCES

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[2] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 530.

[3] Edmond Awad, Sohan Dsouza, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. Crowdsourcing moral machines. *Commun. ACM* 63, 3 (2020), 48–55.

[4] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59.

[5] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. 2018. Using Contextual Bandits with Behavioral Constraints for Constrained Online Movie Recommendation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.* 5802–5804.

[6] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces* (2019).

[7] Ian Bogost. 2018. Enough with the Trolley Problem. *The Atlantic* (2018).

[8] Jonathan Bragg, Mausam Mausam, and Daniel S Weld. 2016. Optimal testing for crowd workers. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 966–974.

[9] Rodney Brooks. 2017. Unexpected Consequences of Self-Driving Cars. *rodneybrooks.com* (2017).

[10] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? (2016).

[11] Yvonne Chen, Travis Mandel, Yun-En Liu, and Zoran Popović. 2016. Crowdsourcing Accurate and Creative Word Problems and Hints. *AAAI HCOMP* (2016).

[12] Albert Mo Kim Cheng, James C Browne, Aloysius K Mok, and R-H Wang. 1991. Estella; a facility for specifying behavioral constraint assertions in real-time rule-based systems. In *COMPASS'91, Proceedings of the Sixth Annual Conference on Computer Assurance*. IEEE, 107–123.

[13] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. 2018. A Lyapunov-based Approach to Safe Reinforcement Learning. *arXiv preprint arXiv:1805.07708* (2018).

[14] Sasha Costanza-Chock. 2018. Design Justice: towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society* (2018).

[15] Peng Dai, Mausam Mausam, and Daniel S Weld. 2011. Artificial intelligence for artificial artificial intelligence. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, 1153–1159.

[16] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. 2018. Safe Exploration in Continuous Action Spaces. *arXiv preprint arXiv:1801.08757* (2018).

[17] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6889–6892.

[18] B. A. Davey and H.A Priestley. 1990. *Introduction to Lattices and Order*. Cambridge University Press.

[19] Greg d'Eon, Joslin Goh, Kate Larson, and Edith Law. 2019. Paying Crowd Workers for Collaborative Work. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[20] Deven R Desai and Joshua A Kroll. 2017. Trust but verify: A guide to algorithms and the law. *Harv. JL & Tech.* 31 (2017), 1.

[21] Giada Di Stefano, Francesca Gino, Gary Pisano, and Bradley Staats. 2014. Learning by Thinking: Overcoming Bias for Action through Reflection. *Harvard Business School Working Paper Series* 58, 14-093 (March 2014).

[22] Brendan Dixon. 2020. The moral machine is bad news for AI ethics. *Mind Matters News* (2020).

[23] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2623–2634.

[24] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. *arXiv preprint arXiv:1711.03938* (2017).

[25] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[26] Jie Gao, Hankz Hankui Zhuo, Subbarao Kambhampati, and Lei Li. 2015. Acquiring Planning Knowledge via Crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

[27] Yotam Gingold, Etienne Vouga, Eitan Grinspun, and Haym Hirsh. 2012. Diamonds from the rough: Improving drawing, painting, and singing via crowdsourcing. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[28] Meghan Holohan. 2018. Her baby was stillborn, but the ads just kept coming: One mother shares her pain. Today, https://www.today.com/parents/gillian-brockell-s-open-letter-tech-companies-goes-viral-t145124.

[29] Bert I Huang. 2019. LAW'S HALO AND THE MORAL MACHINE. *Columbia Law Review* 119, 7 (2019), 1811–1828.

[30] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[31] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020), 169–169.

[32] Walter S Lasecki, Adam Marcus, Jeffrey M Rzeszotarski, and Jeffrey P Bigham. 2014. *Using microtask continuity to improve crowdsourcing*. Technical Report.

[33] Glen D Lawrence. 2013. Dietary fats and health: dietary recommendations in the context of scientific evidence. *Advances in nutrition* 4, 3 (2013), 294–302.

[34] Sang Won Lee, Rebecca Krosnick, Sun Young Park, Brandon Keelean, Sach Vaidya, Stephanie D O'Keefe, and Walter S Lasecki. 2018. Exploring real-time collaboration in crowd-powered systems through a ui design tool. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

[35] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883* (2017).

[36] Tianyi Li, Chandler J Manns, Chris North, and Kurt Luther. 2019. Dropping the baton? Understanding errors and bottlenecks in a crowdsourced sensemaking pipeline. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW

(2019), 1–26.

[37] 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.

[38] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[39] David McAllester and Jeffrey Mark Siskind. 1993. Nondeterministic lisp as a substrate for constraint logic programming. In *Proceedings AAAI*. 133–138.

[40] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[41] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1345–1354.

[42] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[43] David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human computation* 11, 11 (2011).

[44] Barry O'Sullivan. 2002. Interactive constraint-aided conceptual design. *AI EDAM* 16, 4 (2002), 303–328.

[45] 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.

[46] Singer Peter. 1981. The expanding circle: ethics and sociobiology. *New York: Farrar, Straus and Giroux* (1981).

[47] Renee M Petrilli, Gregory D Roach, Drew Dawson, and Nicole Lamond. 2006. The sleep, subjective fatigue, and sustained attention of commercial airline pilots during an international pattern. *Chronobiology international* 23, 6 (2006), 1357–1362.

[48] Niki Pfeifer and Gernot D Kleiter. 2007. Human reasoning with imprecise probabilities: Modus ponens and Denying the antecedent. In *5th International symposium on imprecise probability: Theories and applications*. 347–356.

[49] Anatol Rapoport, Albert M Chammah, and Carol J Orwant. 1965. *Prisoner's dilemma: A study in conflict and cooperation*. Vol. 165. University of Michigan press.

[50] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. 2017. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. *arXiv preprint arXiv:1707.05173* (2017).

[51] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. 2018. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2067–2069.

[52] Stephen Stich. 2007. Evolution, altruism and cognitive architecture: a critique of Sober and Wilson's argument for psychological altruism. *Biology & Philosophy* 22, 2 (2007), 267–281.

[53] Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. MIT press.

[54] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. 2018. Reward Constrained Policy Optimization. *arXiv preprint arXiv:1805.11074* (2018).

[55] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004.

[56] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[57] George Christopher Williams. 1966. *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton university press.

[58] Christine Wolf and Jeanette Blomberg. 2019. Evaluating the Promise of Human-Algorithm Collaborations in Everyday Work Practices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[59] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[60] Hankz Hankui Zhuo. 2015. Crowdsourced Action-Model Acquisition for Planning.. In *AAAI*. 3439–3446.

[61] Martin Zwick and Jeffrey A Fletcher. 2014. Levels of altruism. *Biological Theory* 9, 1 (2014), 100–107.

## A PROOF OF THEOREM 1

Theorem 1. *Aside from the length limit,*[15] *our method of rule construction is fully expressive, that is, it can represent any Boolean function over the set of base predicates.*

Proof. We will show that we can represent any valid Boolean function in disjunctive normal form (DNF). Recall that in first-order logic, each literal in the DNF is composed of a predicate applied to valid arguments.

First, note that our interface directly allows the user to select predicates and their valid arguments. Therefore it is possible to construct all valid literals in our interface.

Now, in the DNF, each literal may be negated by the use of the not operator immediately preceding the literal. Although we do not allow the user to insert an explicit not operator, when choosing predicates we always allow the user to choose its negated form. Therefore, the user may construct any literal or its negation.

We proceed to show that the user can create any sequence of DNF clauses joined by ORs, by **induction on the number of literals**. As part of our inductive hypothesis we also prove an invariant: there is always either exactly one rightmost parenthesis belonging to the current clause[16] after the last literal $D$, or there is zero and $D$ is the only literal in its clause.

Note that in a DNF, the result is the same regardless of the order the ORs are evaluated in, due to the well-known associativity of the OR operator. Therefore it is not necessary to show that the clauses are evaluated left-to-right, just that they are fully evaluated (e.g. by being enclosed in parentheses) and then combined with another clause using the OR operator.

Base Case: The dropdowns allow the user to directly create any single literal. There are no parentheses, but the literal is the only clause member so the inductive hypothesis holds.

Inductive case: Our inductive hypothesis holds for one or more literals and we wish to add an additional literal.

If there is one literal, our interface treats this differently, AND or OR may be added directly to achieve the desired single-clause or two-clause result. In the case of AND, there is a single rightmost parenthesis and multiple elements in the clause, so the invariant holds. In the case of OR, there are no right parentheses belonging to the cause, but the last literal is the lone element of the clause so the invariant holds.

Now, assume there is more than one literal. In that case, there must be at least one logical. Take the rightmost logical. By our method of adding parentheses, it must have added a parenthesis just to the right of the last literal $D$. Further, in order to let the user add the next logical (and, subsequently, literal), our method of construction will add a choicebox around this right parenthesis . We can visually represent this as "…D –) –…'.

Note that, by our inductive hypothesis, either there is exactly one right parenthesis after $D$ belonging to $D$'s clause or there is zero. If there is one, it clearly must be the parenthesis immediately following $D$.[17] Therefore, there are three cases:

(1) The parenthesis just after $D$ does not belong to the clause, and we want to add the current literal F WLOG, into the same clause as the preceding variable D. In this case we know D is the sole element of the clause by the inductive hypothesis. Therefore, we select the inner choicebox and fill it with an AND. Since we chose inner, the left hand parenthesis will go just before D, resulting in "…(D AND F))…". Since by our inductive hypothesis D was correctly

[15] We limit the length of the predicates due to practical storage and data transmission issues.

[16] When we say a parenthesis belongs to the current clause, we mean that it (and it's matching parenthesis) does not encompass any literals outside of the clause.

[17] In order for the parenthesis to the right of the parenthesis after $D$ to belong to $D$'s clause, the inner parenthesis just after $D$ would have to as well.

in a clause by itself before adding $F$, clearly $D$ and $F$ are now correctly in a clause together, as desired. We have multiple members in a clause and exactly one right parenthesis belonging to the clause to the right of $F$, so the invariant holds.

(2) The parenthesis just after $D$ belongs to the clause, and we want to add the current literal F WLOG, into the same clause as the preceding variable D. In this case one selects the outer choicebox, resulting in "…D) AND F)…". Since the parenthesis just after $D$ belongs to the clause, the matching left parenthesis must be contained within the clause. The left hand parenthesis will go just before D's left parenthesis, therefore since D was correctly a part of the clause by the inductive hypothesis, F must be as well. We have multiple members in a clause and exactly one right parenthesis belonging to the clause to the right of $F$, so the invariant holds.

(3) We want F to be the first element of a new clause. We select the outer choicebox and fill it with an OR, giving us "…D ) OR F)…". Now, as previously shown, because of the induction hypothesis the right parenthesis immediately following D is the only right parenthesis after D that can possibly belong to the clause. Therefore, OR and F must correctly be outside of $D$'s clause. Further, by the inductive hypothesis the DNF expression was built correctly thus far, so if the OR and F are outside of $D$'s clause they must also be outside all clauses.[18] The only literal in the clause is F, and there are no right parentheses belonging to the clause, so the invariant holds.

□

Received January 2020; revised June 2020; accepted July 2020

[18] Or they would be in a clause that somehow incorrectly has D's clause nested inside of it, violating the inductive hypothesis.