Belief Miner: A Methodology for Discovering Causal Beliefs and Causal Illusions from General Populations

SHAHREEN SALIM, Stony Brook University, USA
MD NAIMUL HOQUE, University of Maryland, USA
KLAUS MUELLER, Stony Brook University, USA

Causal belief is a cognitive practice that humans apply everyday to reason about cause and effect relations between factors, phenomena, or events. Like optical illusions, humans are prone to drawing causal relations between events that are only coincidental (i.e., causal illusions). Researchers in domains such as cognitive psychology and healthcare often use logistically expensive experiments to understand causal beliefs and illusions. In this paper, we propose **Belief Miner**, a crowdsourcing method for evaluating people's causal beliefs and illusions. Our method uses the (dis)similarities between the causal relations collected from the crowds and experts to surface the causal beliefs and illusions. Through an iterative design process, we developed a web-based interface for collecting causal relations from a target population. We then conducted a crowdsourced experiment with 101 workers on Amazon Mechanical Turk and Prolific using this interface and analyzed the collected data with Belief Miner. We discovered a variety of causal beliefs and potential illusions, and we report the design implications for future research.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**.

Additional Key Words and Phrases: Causal Beliefs, Causal Illusion, Crowdsourcing, Evaluation Method

ACM Reference Format:
Shahreen Salim, Md Naimul Hoque, and Klaus Mueller. 2024. Belief Miner: A Methodology for Discovering Causal Beliefs and Causal Illusions from General Populations. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 21 (April 2024), 37 pages. https://doi.org/10.1145/3637298

1 INTRODUCTION

In the view of psychology, a belief is "the mental acceptance or conviction in the truth or actuality of some idea" [68]. Accordingly, a *causal* belief is a belief about one or more factors that are thought to cause or contribute to the development of a certain phenomenon, such as an illness or the outcome of an intervention. A causal belief is not restricted to a single relation; it can embrace entire causal mechanisms. As Blanzieri [10] writes, causal beliefs are halfway between actual knowledge about a physically objective reality and a socially-constructed reality. They are different from causality and causal inference which are strictly derived from hard data via well-defined statistical principles [10].

An interesting phenomenon is that of *causal illusion* which occurs when people develop the belief that there is a causal connection between two events that are in fact just coincidental [54]. Causal illusions are the underpinnings of pseudoscience and superstitious thinking and they can have disastrous consequences in many critical areas such as health, finances, and well-being [29].

Authors' addresses: Shahreen Salim, ssalimaunti@cs.stonybrook.edu, Stony Brook University, Engineering Drive, Stony Brook, New York, USA, 11790; Md Naimul Hoque, University of Maryland, 4130 Campus Dr, College Park, Maryland, USA, nhoque@umd.edu; Klaus Mueller, Stony Brook University, Engineering Drive, Stony Brook, New York, USA, 11790, mueller@cs.stonybrook.edu.

ACM 2573-0142/2024/4-ART21
https://doi.org/10.1145/3637298

A mild form of a causal illusion is the good-luck charm that people carry to special events or in general, but more serious are bogus medicines that can inhibit people from taking up scientifically more credible treatments to restore or preserve their health.

The first step towards combating a causal illusion is to detect it. Researchers typically use *contingency judgement tasks* for this purpose [54]. In these experiments, participants are exposed to a number of trials (around 50) in which a given cause is present or absent, followed by the presence or absence of a potential outcome. At the end of the experiment, participants are asked to judge the causal relations between the cause and the effect (or outcome). While effective, these experiments have several limitations: 1) they lack a mechanism to expose complex causal structures (with many links and chains) beyond a single cause and effect; 2) they are unable to elicit complex cognitive conflicts such as how people fare with competing causes to the same effect; and 3) they are highly controlled laboratory studies and logistically expensive to conduct. We believe that crowd-sourcing these activities can provide a powerful alternative tool to address these challenges.

We propose **Belief Miner**, a crowdsourcing methodology for discovering causal beliefs and illusions from the general population. While prior works have showcased the capability of crowdsourcing in identifying causal relations [5, 82, 83], they often approach the topic from a predominantly data-driven perspective, focusing either on creating large networks or training datasets for Causal ML. Prior works were not designed to model contingency judgment tasks and identify causal illusions using crowdsourcing. "Belief Miner" aims to fill this gap.

Our methodology draws on the rich literature in psychology, causal illusion, and crowdsourcing (Section 2, 3, and 4). Informed by the literature, our method uses crowdsourcing to collect a dataset of causal relations from a general population on a topic of interest and contrasts these with causal relations obtained from domain experts to understand causal beliefs and detect illusions via several metrics and mechanisms we propose. Thus, Belief Miner also offers a nuanced and formalized post-hoc analysis [5] in causal crowdsourcing, which current studies and systems lack [5, 82, 83].

To validate our method, we designed an interactive web-based tool that would allow crowd workers to interactively create small causal networks (and alter their created network if needed) from a set of variables. We then conducted a formative study with 94 crowd workers on Amazon Mechanical Turk (AMT). We asked participants to create causal relations between randomly chosen variables relevant to climate change. We selected the theme of climate change due to its propensity for controversial views and potential to expose causal illusions. For instance, in a study conducted in 2019 [48], a slight majority (59%) of Americans believed climate change is human-caused, and nearly a third (30%) believed that natural variability is the primary cause. This belief and confusion can be harmful to necessary policy-making to counter climate change [27].

As mentioned, in addition to the crowdsourced data, we also collected causal relations from a group of domain experts (e.g., climate scientists). Then, with both datasets in place, we employed our method on them to discover causal illusions. While the results were generally positive, we identified two issues: 1) there was a moderate possibility of selection bias due to the order in which the variables were presented to the participants; and 2) the completion time was longer than expected which pointed to possible usability issues.

Based on the findings of this formative study, we revised the design of the interface and the experimental protocol. Using the revised design, we conducted another study with 101 crowd workers from AMT and Prolific. We observed a stronger alignment between the causal beliefs of the crowd and the experts and a reduced completion time of the crowdworkers compared to the formative study. Our findings also reveal various discrepancies between the causal relations created by crowd workers and experts. We observe 1) that a significant number of workers overestimated the impact of certain attributes, and 2) that participants with flawed causal beliefs (i.e., illusions)

	Outcome present	Outcome not present
Cause present	$O\|C$	$\neg O\|C$
Cause not present	$O\|\neg C$	$\neg O\|\neg C$

Table 1. **Contingency table components**. In this table, C is the cause and O is the outcome.

	Trial Matrix 1			Trial Matrix 2		
	Outcome present	**Outcome not present**	**Probabilities**	**Outcome present**	**Outcome not present**	**Probabilities**
Cause present	80	20	$P(O\|C) = 0.8$	20	80	$P(O\|C) = 0.2$
Cause not present	20	80	$P(O\|\neg C) = 0.2$	80	20	$P(O\|\neg C) = 0.8$
			$\Delta p = 0.6$			$\Delta p = -0.6$
	Trial Matrix 3			**Trial Matrix 4**		
	Outcome present	**Outcome not present**	**Probabilities**	**Outcome present**	**Outcome not present**	**Probabilities**
Cause present	80	20	$P(O\|C) = 0.8$	20	80	$P(O\|C) = 0.2$
Cause not present	80	20	$P(O\|\neg C) = 0.8$	20	80	$P(O\|\neg C) = 0.2$
			$\Delta p = 0.0$			$\Delta p = 0.0$

Table 2. **Trial matrices emerging from different contingency table configurations**. Each quadrant is an example of one of the four types of trial matrices. The two matrices on the bottom represent null contingencies. All numbers are in %.

assigned lower confidence scores to their networks in both studies, suggesting that our mechanism effectively counters illusion and has the potential to increase awareness among individuals.

In summary, our contributions are as follows: 1) *Belief Miner*, a methodology that includes a web-based interactive system and evaluation method for discovering causal beliefs and illusions; and 2) Two crowdsourcing studies on Amazon Mechanical Turk and Prolific with 94 and 101 crowd workers. The collected data from the experiment shows an application of our methodology in the domain of climate change.

2 BACKGROUND: CAUSAL BELIEF AND CAUSAL ILLUSION

In this section, we will provide the background that has guided our research and development. We will build on a principal metric, the Δp index [3], which researchers use to design the level of contingency in an experiment.

In the simplest case, there is one potential cause C and one observed outcome O. Given these two variables there are then four possible configurations: 1) both C and O occur, 2) C occurs but O may not, 3) C may not occur but O still does, and 4) neither C nor O occur. We can capture these four configurations into the contingency table shown in Table 1. Δp is defined as follows [3, 54]:

$$\Delta p = P(O|C) - P(O|\neg C)$$
$$P(O|C) = \|O|C\|/(\|O|C\| + \|\neg O|C\|)$$
$$P(O|\neg C) = \|O|\neg C\|/(\|O|\neg C\| + \|\neg O|\neg C\|)$$

A true causal relationship exists when Δp is non-zero. When Δp is positive then C is said to promote O, while when Δp is negative C is said to inhibit O.

The contingency table gives rise to four distinct cases of trial matrices shown in Table 2. In the first case, $\Delta p > 0$, C might be an evidence-based medicine to treat a cold and O is the disappearance of the cold. This is shown in trial matrix 1 where row 1 is the treatment that divides the response of a cohort of patients who took the medicine, while row 2 is the control that divides the response of a cohort of patients who did not take the medicine and instead took a placebo.

In the second case, $\Delta p < 0$, C might be an effective medicine to control a person's cholesterol level, and $\neg O$ is the cholesterol level that remains in check and will not rise. This case is illustrated in trial matrix 2 which is trial matrix 1 transposed.

Trial matrix 3 is one for which $\Delta p = 0$, which occurs when $P(O|C) = P(O|\neg C)$ — the null contingency. In this case, the treatment has no effect on the outcome. A patient may take the cold medicine or not, but the cold will always disappear on its own. A classic example of this scenario is alternative (homeopathic) medicine which typically lacks strong scientific evidence for its effectiveness [33, 70].

The symmetric case of trial matrix 4 is analogous. A person with no risk of high cholesterol might take, inspired by effective product marketing, a scientifically unproven medicine and feel confirmed in that choice when the level stays normal.

2.1 When a Causal Illusion Weakens Trust in a Proven Cause

Having formed a null contingency about a certain phenomenon can lead people to discount a true causal relationship of a scientifically acknowledged cause for the same outcome. As Matute et al. [54] write: "The availability of more than one potential cause can result in a competition between both causes so that if one is considered to be a strong candidate, the other will be seen as a weak one." This was verified via several experiments where it was found that participants who had established a prior belief about the effectiveness of an unproven *bogus* medicine – a causal illusion – weakened the belief in the effectiveness of a proven medicine. In contrast, participants who did not have a chance to develop the illusion had sustained trust in the proven medicine [73].

Experiments have shown that the strength of a causal illusion can be effectively controlled by the frequency at which the cause is present, even when its effectiveness to drive the outcome remains the same. This has important implications on a person's belief in a proven medicine. The more frequently the competing (alternative) medicine has been administered and a confirmatory (yet specious) outcome has been experienced, either now or in the past, the smaller the belief in the proven medicine [80].

An effective way to convince people that they have fallen victim to the null contingency is to ask them not to take the treatment when they hope for an outcome to occur [8]. But this proves difficult when the cost of the treatment is low and the outcome is ubiquitous and persuasive. In fact, the ubiquity of both treatment and outcome are perfect conditions for a causal illusion to emerge [74, 81]. It fosters trust in the belief that there must be a causal relationship between the two since there are many opportunities for coincidences [9].

While the examples given so far are relatively benign, there are more serious scenarios where these cognitive mechanisms can be harmful [29]. A person with a natural medicine mindset, when

receiving, say, a cancer diagnosis might resort to acupuncture, herbal treatments, fruit juice therapy, and spiritual consultations instead of seeking more conventional evidence-based interventions, such as surgery or radiation, and chemotherapy. This is what has been reported to have happened to Steve Jobs, the founder of Apple Computer and an exceptionally tech-savvy and forward-thinking individual [13]. It vividly shows that gaining immunity from null contingencies is hard.

2.2 Gauging Causal Illusions: From the Lab into the Wild

Causal illusion has been of interest to researchers for a long time. It can reveal many cognitive and behavioral practices, as described in Section 2.1. To study different research questions, researchers typically tweak the trial matrices (Table 2) to create desired scenarios and then conduct control experiments based on that.

These experiments typically involve hypothetical scenarios set in the medical domain. Participants are asked to impersonate doctors who are assigned a set of fictitious computer-modeled patients (i.e., the total number of trials). In some scenarios, the participants take on an active role in prescribing medical treatments for some illnesses, while in others they simply observe the patients follow a certain treatment regime. As the experiment proceeds, participants see records of patients who either have or have not taken a certain medicine and then have or have not recovered from the disease (based on the trial matrices). At the end of the experiment, the participants are asked whether the treatment was effective or not (for example see [54]).

The experiments are meticulously designed to expose the triggers and impacts of causal illusions. They are highly controlled laboratory studies and focus on very simple and elementary causal relationships. Our crowd-based belief miner takes these studies out of the research laboratory into the wild where an abundance of data awaits, reflecting complex causal chains with many links and paths. There are also many unexpected cause and outcome variables that might be discovered when the crowd-sourced data are studied in depth. It is a tool by which complex cognitive conflicts can be efficiently extracted and exposed for real-life phenomena of possibly high complexity. While this approach cannot directly replicate the usage of trial matrices shown in Table 2, it can produce the conclusions that are typically derived from the experiments that use trial matrices. Furthermore, it is easy to set up for practitioners.

Another interesting aspect of our methodology is that it asks participants to engage in a critical assessment of their beliefs, by ways of reviewing the small causal network they construct. As we have already hinted at in the introduction, we found evidence of the implications of a theory proposed by Walsh and Sloman [75], who experimented with the concept of contradiction as a way to get people to revise their beliefs in a causal illusion. They suggest that a person might reduce their belief in the effectiveness of a certain treatment upon discovering that some elements of it have an outcome that is opposite of what they expected.

3 RELATED WORK

Our goal in this paper is to discover causal beliefs and illusions using crowdsourcing. In this section, we discuss crowdsourcing, HCI, and CSCW concepts relevant to achieve that.

3.1 Crowdsourcing and Quality Control in Crowdsourcing

Bigham et al. [7] identified three broad areas for collective intelligence and HCI. They are 1) *directed crowdsourcing*, where a single person or a group guides a large set of people to accomplish a task (e.g., labeling a large dataset [15, 37], CommunityCrit [53]); 2) *collaborative crowdsourcing*, where a group of people works together to accomplish a task (e.g., Wikipedia, Project Sidewalk [67], ConceptScape [50]); and 3) finally, *passive crowdsourcing*, where people do not coordinate and are

not consciously aware of participating in a crowdsourced system, however, one can still mine their behavior to infer collective intelligence (e.g., mining search history).

In this paper, we mainly focus on directed crowdsourcing (referred to as crowdsourcing from here on) since we provided explicit direction to our crowd workers. In a crowdsourcing experiment, a crowd worker typically completes a small part of the overall task (i.e., micro-tasks) [19, 40, 43]. These micro-tasks are eventually aggregated for inferring collective intelligence. In our case, a micro-task refers to creating a causal network between a small number of attributes.

Over the years, several methods and tools have been proposed for measuring and increasing the efficacy of the design of the micro-tasks based crowdsourcing [6, 41–43, 57, 66]. Of particular interest is *Quality Control* in crowdsourcing [42], which ensures the validity of the collected response from crowd workers. This is important since crowd workers tend to spend minimum effort and sometimes try to game the systems [42]. The currently established methods for quality controls typically fall into two broad categories: task design and post-hoc analysis [42]. Examples for task design include fault-tolerant subtasks [6, 43, 44, 57], attention check [1], peer review filters [6, 23, 35, 43], intelligent task assignment [39] and optimizing instructions [24, 40, 47, 63]. Examples for post-hoc analysis include comparison with gold standards [5, 12, 24], validation study [67], agreement between crowd workers [12, 36], and behavior analysis [16, 30, 65, 66, 84].

Our work can be seen as a post-hoc analysis model for causal crowdsourcing. We use causal relations collected from experts (i.e., gold standards), a popular post-hoc method, to find illusions in the crowd-generated causal networks. Although relevant work in causal crowdsourcing [5] employed comparison with gold standards on a smaller scale, our work both scales up the approach and adds a nuanced dimension to the post-hoc analysis model tailored for causal crowdsourcing. This makes our contribution one of the first to delve into such depth and granularity in this domain. Finally, we believe our findings will guide future task designs for causal crowdsourcing. Appropriate task designs can make people self-aware and help them avoid falling victim to causal illusions. We lay down this future direction in our discussion (Section 11).

3.2 Causality and Crowdsourcing

While causality is a core concept across several scientific domains, designing crowdsourced experiments for collecting causal relations is a relatively new research area. Caselli et al. contributed to this area by focusing on annotating causal relations in narrative texts, specifically news data, through crowdsourcing experiments [14]. Their work analyzed parameters affecting annotation quality and compared crowdsourced and expert annotations, emphasizing the generation of structured data based on narrative strategies. The most relevant work in this space, however, is Iterative Pathway Refinement [5], a network search strategy where workers modify a short linear pathway between attributes. The authors showed that their method is more efficient than a single line-based micro task, provides better contexts to crowd workers, and the union of the pathways can create a large network. Yen et al. [82] extended this line of work by proposing CausalIDEA, an interactive interface where users can create small networks as well as provide textual explanations for creating specific causal relations. The causal diagrams and textual narratives add a new dimension to understanding causal beliefs. The authors investigated how a user's causal perception is affected by seeing the causal networks created by others. Furthermore, Yen et al. introduced CrowdIDEA [83], a tool that integrates crowd intelligence and data analytics to support causal reasoning, featuring a three-panel setup: enabling access to crowd's causal beliefs, data analytics, and the ability to draw causal diagrams. Their study also demonstrated that seeing the crowd's causal beliefs significantly improved the accuracy of causal relationships in the final diagrams of the participants and reduced reliance on data analytics, showcasing the tool's potential to enhance causal reasoning processes.

These prior works provide evidence that people can identify causal relations between pairs of events and collaboratively create causal networks that are information-rich. While inspiring, they consider causal crowdsourcing from an algorithmic or data science perspective, focusing either on creating large networks or training datasets for Causal ML [45]. In contrast, we consider causal crowdsourcing as a tool to surface how causal illusions persist in a domain, which prior works have largely overlooked [5, 82].

While we do not focus on data science or machine learning, our work has implications for Causal ML. For example, failing to detect causal illusions and erroneous causal relations can introduce biases into datasets, leading to skewed decision-making based on inferred causal relations [31, 34, 76]. Broader ML literature has also demonstrated that ML models trained with datasets containing spurious relations or errors may perform poorly and yield incorrect inferences [45, 60, 64, 79].

In summary, the absence of mechanisms to detect causal illusions is a significant obstacle to the practical application of causal crowdsourcing in different domains, including social science and policy-making (domain of this work), causal ML, finances, and health. We aim to bridge this gap.

4 METHOD

In this section, we describe the evaluation method we have devised to help expose the potential causal illusions and the complex cognitive conflicts mentioned in Section 2.

4.1 Data

4.1.1 Causal Belief Data. Causal beliefs are conceptualized through the presence of a *cause* and an *effect* and their *relationship* in a certain phenomenon. Following the Structural Causal Model (SCM) [61], we define a *causal relationship* through a directed link/edge between the cause and effect (*cause* $\rightarrow$ *effect*). Therefore, gathering causal relations from a domain of interest is synonymous with identifying the causal edges among a collection of relevant attributes that form an interconnected directed causal network [5].

Let us define a causal network $N = (V, E)$ with nodes V and links E. Here, V is a set of causal attributes and $E \subseteq \{u \rightarrow v | u, v \in V \textit{ and } u \neq v\}$, where $u \rightarrow v$ is a unidirectional causal relationship between attributes u and v.

4.1.2 Ground Truth. Comparison with gold standards or ground truth is common in crowdsourcing [5, 12, 24]. In our case, ground truth refers to scientifically verified causal relations. However, establishing ground truth in a complex domain such as climate change where many interconnected factors may exist, can be a demanding task. We propose the following collaborative method with two *phases* to establish ground truth for all possible causal links.

Phase1: Independent Ground Truth Generation. First, a fixed number of experts are instructed to create their version of the causal networks independently using all the attributes. We propose a minimum of three experts. We require experts to provide scientifically verified references to the created relations.

Phase2: Collaborative Meeting. After the network creation phase, the union of their networks is considered for the discussion phase. Here, the experts work collaboratively to assign *credibility scores* for all possible causal links in the unified network. Credibility scores denote the level of validity for a specific causal link. We chose this notion since there can be "levels" of correctness for a specific causal relation instead of them being just right or wrong. This can also be described as the strength of a causal relationship. The notion is motivated by mediation analysis in causality [52]. According to mediation analysis [52], in a causal network, an attribute may have a direct effect on another attribute as well as indirect effects through other attributes (mediating attributes). Thus,

the number of mediating attributes between two attributes indicates the strength of the causal relationship.

Based on this observation, we propose four levels for the credibility scores: all links that did not appear in any experts' causal network can be assigned the lowest possible credibility score (0), and the links that appear in all of them can be assigned the highest credibility score (3). Finally, the rest of the links are assigned a score of 1 or 2 after expert deliberation, depending on the number of mediating variables present in between.

4.2 Metrics

We offer two angles of evaluation or analysis of the collected causal belief data: 1) an aggregated quantitative and qualitative overview of the causal beliefs; and 2) causal illusion detection. The metrics required are defined below:

4.2.1 Aggregated Evaluation: Aggregated statistics such as the distribution of total votes are common in crowdsourcing [42], including causal crowdsourcing [5, 82]. They are useful for obtaining an overview of the data and determining outliers. In addition to the total votes, we also examine the distribution for the *Average Network Credibility Score (ANC)*. We calculate the average network credibility (ANC) score using the following equation:

$$ANC_N = \frac{\sum_{e \in E} cs_e}{|E|}$$

Here, N is a small causal network created by an individual, and cs_e is the credibility score (from Section 4.1.2) of link e present in network N. Thus, ANC_N can be calculated for each network created by separate individuals, and their distribution will indicate the credibility for the networks created by people.

4.2.2 Causal Illusion Detection. In Section 2, we define *Causal Illusion* as the incorrect assumption of cause and effect in a certain phenomenon. Thus, causal illusion inherently denotes a discrepancy between people's causal beliefs and the ground truth.

We define two types of causal illusions that utilize ground truth data and the causal belief data collected from the people/crowd. We represent the crowd data using the *crowd score (cr)* (a score representing the crowd's inclination toward a specific causal link) and the ground truth data using the *credibility score (cs)*. There are several ways the *crowd score (cr)* can be calculated, such as using the normalized total votes assigned to a causal link by the crowd.

Let us suppose we have a causal link l with a crowd score of cr_l and a credibility score (from ground truth) of cs_l. There can be two potential cases of causal illusion;

- Where the crowd had a stronger inclination toward a causal link with a comparatively lower credibility score, i.e., $cr_l > cs_l$. We define this state as being *potentially misinformed*. While we do not investigate the reasons behind this, they can indicate the consumption of less credible sources, shallow reading practices, and denial of climate change [71].
- Where the crowd had a weaker inclination toward a causal link with a comparatively higher credibility score, i.e., $cr_l < cs_l$. We define this state as the state of being *potentially uninformed/oblivious*. The cause behind obliviousness can be a lack of knowledge regarding that specific topic, i.e., people genuinely do not know about it enough to vote for it.

5 INITIAL INTERFACE

Our overarching goal in designing the interface was to foster critical thinking about a topic of interest (e.g., climate change) while ensuring ease of use in creating causal networks. We felt the necessity to develop our own data collection tool since there is no open-source tool available for

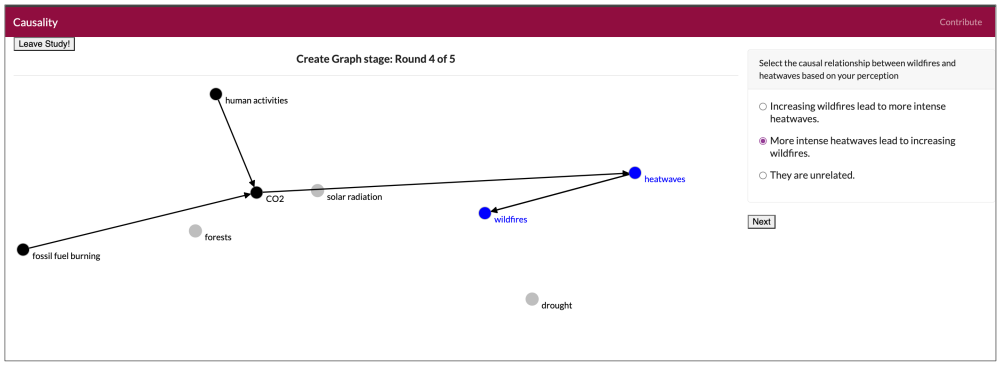

Fig. 1. **Overview of the initial collection interface**. This example shows the steps participants followed to create the causal networks. In the center, we see the causal network created by the participant. The edge connecting the two nodes marked in blue is the newly created edge by the participant. We provide options to select the direction for the newly created causal edge on the right.

causal belief collection. In pursuit of this objective, we employed an iterative design process with specific design goals to continually refine and improve the interface based on evolving needs.

5.1 Design Goals

In this section, we provide the initial design goals for the interface. Some part of the interface loosely follows the methodology proposed in prior works [5, 82]: people create small causal networks to demonstrate their beliefs, and then the small networks are aggregated into a large causal network. However, we non-trivially enhanced the methodology to meet the following design goals:

DG1. Interactively Create and Modify Causal Networks. Interactive visual interfaces can enhance people's understanding and decision-making of complex systems [22]. We decided to utilize an interactive visual interface based on a Directed Acyclic Graph (DAG) to represent causal networks, enabling users to freely choose attributes and their causal relations. The interface should provide modification controls, such as changing link directions, to allow users to refine networks.

DG2. Use of Natural Language to Narrate Causality. Natural language texts are used in conjunction with visual representations to provide explanations and enhance comprehension of causal networks [82]. This inspired us to use natural language as an explanation along with the interactive visual representations [17]. The narrative component should clarify potential confusion arising from graphical representations.

DG3. Quick Completion Time. One important design goal in crowdsourcing studies is to ensure quick completion time to enhance the efficiency of data collection [42]. By minimizing the time required for participants to create, modify, and evaluate causal networks, we want to harness the crowd's collective intelligence more effectively and gather a larger volume of diverse causal networks for analysis and insights.

5.2 Processing Raw Data and Generating Causal Attributes

Before collecting causal beliefs, the set of single/multi-word attributes relevant to the domain must be identified. Experiments with techniques purposed to automatically extract attributes specific to "climate change" (our demonstration domain) from relevant text documents were only mildly successful. To identify the relevant attributes for our domain (climate change), we manually

extracted them from reputable climate-related sources [1]. We also determined specific words to represent upward and downward trends, ensuring natural comprehension, such as "fewer (human activities)", "less (methane)", and "decreasing (solar radiation)". This process yielded a table of 17 attributes with trend terms (34 attributes in total, combining trends with attributes). We then validated these attributes via an informal discussion session with three climate science experts from our university. All experts hold PhDs in relevant fields and have been conducting climate science research for at least ten years. All agreed that the attributes are of importance to any climate science expert, and understanding people's perceptions about them is crucial.

In addition to extracting the attributes, we also use Word2Vec [56], a neural word embedding model trained on the English Wikipedia corpus containing many climate-related documents, to compute attribute coordinates based on semantic distances. We used these coordinates to lay out the variables in the 2D space of the initial interface (Section 5.3). The Word Mover Distance (WMD) was employed for multi-word attributes, calculating the minimum distance words need to travel between documents. We also explored other text embedding models, including BERT [21], RoBERTa [51], and GloVe [62]. After a discussion with the research team, we found Word2Vec's embeddings to be the most suitable option.

We note that we did not utilize the Word2Vec coordinates in our final and redesigned interface because we adopted a different layout approach (Section 8.1).

5.3 Collection Interface Modules

The collection interface is a web-based interface implemented using Python as the back-end language and D3 for visualization [11]. We used MongoDB as a database for our collected results. The input to the interface is a dataset containing nodes of a causal network (V) where the edge list (E) is unknown. Thus, this interface is our tool to infer E from the crowd. The detail of the input dataset is provided in Section 5.2. We describe the visual component of the several interface modules in the following sections. The modules are independent and can be implemented according to the intended sequential workflow. Snapshots of each module are provided as supplemental material. We further describe our workflow specific to the experimental setup later in Section 6.

5.3.1 Instructions and Overview Module. This module provides an overview of the interface and study tasks as a step-by-step guide, with necessary explanations and instructions for each step. It also presents necessary pictures of each page that people will encounter in their workflow.

5.3.2 Demographics Survey Module. This module collects participants' demographics and their perception of domain-specific (e.g., climate change) knowledge and awareness. We currently support multiple-choice and Likert scale questions in this module.

5.3.3 Causal Network Creation Module. The main module in the tool allows participants to create causal relations between pairs of attributes to build a small causal network **(DG1)**. The module follows **DG1** and **DG2** principles, visualizing causal links as node connections and describing them in natural language. The attributes are presented as circular nodes in the interface, positioned based on their word-vector space. The order of the attributes' appearances is pre-determined. Therefore, all individuals will see the attributes in the same order. We did this to observe the difference in people's perceptions given the same set of choices. This makes people's perception a random variable in our experiment instead of the order of the attribute's appearance.

[1] https://www.climaterealityproject.org/blog/key-terms-you-need-understand-climate-change
https://www.climaterealityproject.org/blog/10-indicators-that-show-climate-change
https://opr.ca.gov/facts/common-denier-arguments.html

Participants perform two micro-tasks per causal link: choosing trends for attributes (e.g., increasing and decreasing for the attribute CO2) and selecting the causal relationship between them. This process is repeated in multiple rounds to create the network (Figure 1). Following [5], we decided to keep a narrative flow in the causal network. Therefore, after the first round, a participant needs to create a causal link between any previously chosen attribute and one new attribute. Participants need to make three different causal networks. Except for the first network, participants are presented with a mix of previously used and new attributes when creating a causal network. This ensures reduced learning requirements, which contributes to shorter completion times **(DG3)**.

5.3.4 Alteration Module. Following **DG1**, we developed this module to allow people to alter the network they have created by modifying each network link. The available options are (1) changing the originally selected link direction or (2) deleting the link entirely. This module always appears after the *Creation Module*.

5.3.5 Interpretation and Evaluation Module. In this final step, each individual is asked to evaluate their created (and possibly altered) network. Following **DG2**, we provided a narration of the network. To generate this narration, we combine a graph traversal algorithm with a reasonably simple text template to translate the network into a textual narration. The module then provides people with the opportunity to evaluate their created network. The evaluation is collected as a confidence level on a 5-point Likert scale. This module always appears after the *Alteration Module*.

5.3.6 Usability Rating Module. The purpose of this module is to provide participants with the opportunity to evaluate the complete data collection interface. Following the System Usability Scale (SUS) [72], we present each individual with five usability-related statements. We also provide them with two knowledge-related statements to measure their evaluation of the interface from the perspective of gaining knowledge. A primary goal of the knowledge/learning-related statements was to gauge active thinking's effects by creating a causal network on people's original perceptions. The positive and negative usability statements are presented in alternating order. We mention the seven statements as supplemental materials.

6 FORMATIVE STUDY

We recruited 98 crowd workers from Amazon Mechanical Turk (AMT) to collect causal perceptions on climate change. They used the initial collection interface to create small causal networks. We include all study materials in the supplement.

Crowd Workers' Demographics and Expertise Level Regarding Climate Change. We needed to discard the work of 4 workers due to incompleteness, which led us to have 94 valid workers. We present various aspects of the crowd workers' demographics in Figure 2. The majority of the crowd workers happened to be male, white, and within the age group of 20-40. More than half of the crowd workers finished their bachelor's degrees, and more than two-thirds are employed for wages. Geographically, we only collected results from the United States. A significant portion of the crowd workers is from the Southern region of the United States.

We also examine the self-reported knowledge and agreement levels of crowd workers regarding climate change-related attributes and statements. Around 49% the of participants consider themselves knowledgeable about climate change attributes. Additionally, there is a strong agreement among participants with climate change-related statements (around 90%), indicating a belief in climate change. Further details can be found in Appendix A.1.

Protocol. We used the "external HIT" function on AMT, where the interface hosted on our server was accessible to the workers. Therefore, our interface did not require the crowd workers to log

Fig. 2. **Demographics of the crowd workers in the formative study.** Y-axes represent counts for each category.

in and provide personal information beforehand. We only accepted the results when the crowd workers completed every task in the workflow. The successful crowd workers were paid $2.75 each upon completion. We initially estimated the work would take around 15 minutes. However, the average time taken by the workers was around 30 minutes, according to the AMT website. We provide the step-wise protocol below:

(1) **Read the instructions and pass the test**: In the "Instructions and Overview Module" (Section 5.3.1), the crowd workers received a step-by-step guide and explanations of each interface module. They had the option to revisit previous pages and restart if needed. A test required them to demonstrate their understanding, and they could retry it to improve their comprehension.
(2) **Complete the demographics survey**: In the "Demographics Survey Module" (Section 5.3.2), the crowd workers answered 11 questions about their ethnicity, gender, marital status, geographical location (state), education, employment status, age group, knowledge, concern, and agreement towards climate change. They had the option not to provide their information for each demographic question.
(3) **Create a causal network**: The crowd workers used the "Causal Network Creation Module" (Section 5.3.3) to build a small causal network. They created five causal links by (i) selecting two attributes along with their trends (e.g., "CO2" with an "increasing" or "decreasing" trend) and (ii) choosing the causal relationship between the selected attributes (e.g., *increasing emissions* leads to *increasing CO2*). This process involved two micro-tasks for each link creation. After the first two links, the workers are asked to select a new node that has not been selected before and an already selected node in order to create a new causal relation, adding to the emerging small network.
(4) **Alter causal network**: In the "Alteration Module" (Section 5.3.4), the crowd workers could modify their created causal network. They selected a link, left-clicked on it, and chose from available options for alteration, including *deletion*, *changing the direction*, or *no modification*.

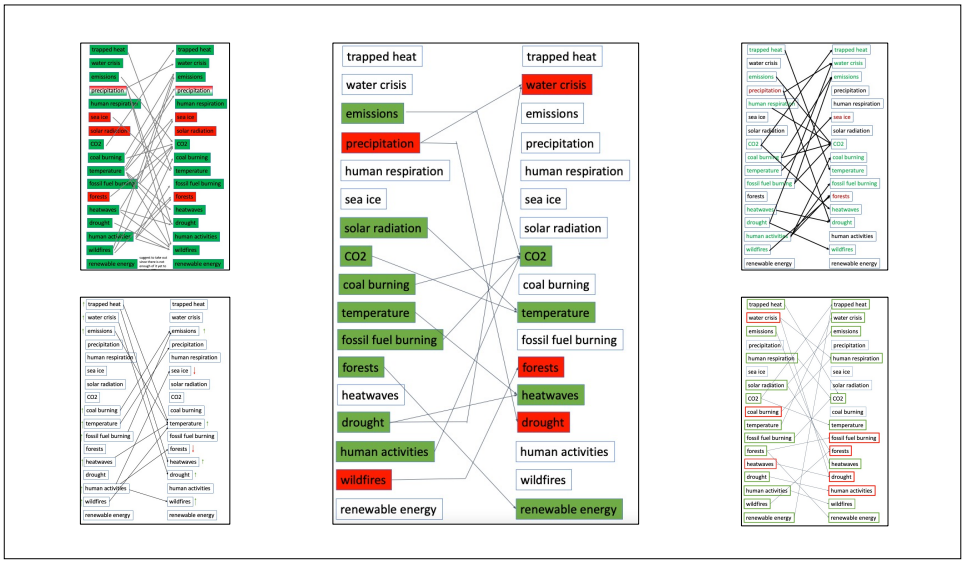

Fig. 3. **Snippets of the ground truth networks created by the experts**. Each expert was provided with the same template containing all causal attributes. Both the left and the right columns contain the same attributes. The experts colored green and red for expressing the upward and downward trend respectively and connected any two attributes of choice from the left column to the right column using an arrow (→) to create a causal link.

(5) **Evaluate causal network**: Crowd workers used the "Interpretation and Evaluation Module" (Section 5.3.5) to review and evaluate their created causal network. They could view it in a node-link diagram or Directed Acyclic Graph (DAG) format and read it in natural language text. Additionally, they provided a confidence level for each network on a scale of 1 to 5.
(6) **Repeat Task 3-5 two more times**: Crowd workers repeated the process of creating small networks, altering them, and evaluating them two more times. The interface provided necessary prompts and repetition for this task.
(7) **Evaluate the interface**: After creating three causal networks, the crowd workers used the "Usability Rating Module" (Section 5.3.6) to rate the interface. They rated seven usability and learning statements on a 5-point Likert scale to provide feedback.
(8) **Verification and compensation**: At the end of their participation, each crowd worker received a unique code. This code was used for result validation, filtering incomplete data, and providing compensation for their contribution.

7 FINDINGS FROM THE FORMATIVE STUDY

We collected three types of data in this experiment: demographic information (94 users), small causal networks (282 networks), and subjective feedback (94 users). The causal networks were saved as vectors containing causal links or (source node, target node) pairs, along with confidence scores tied to the worker identifier.

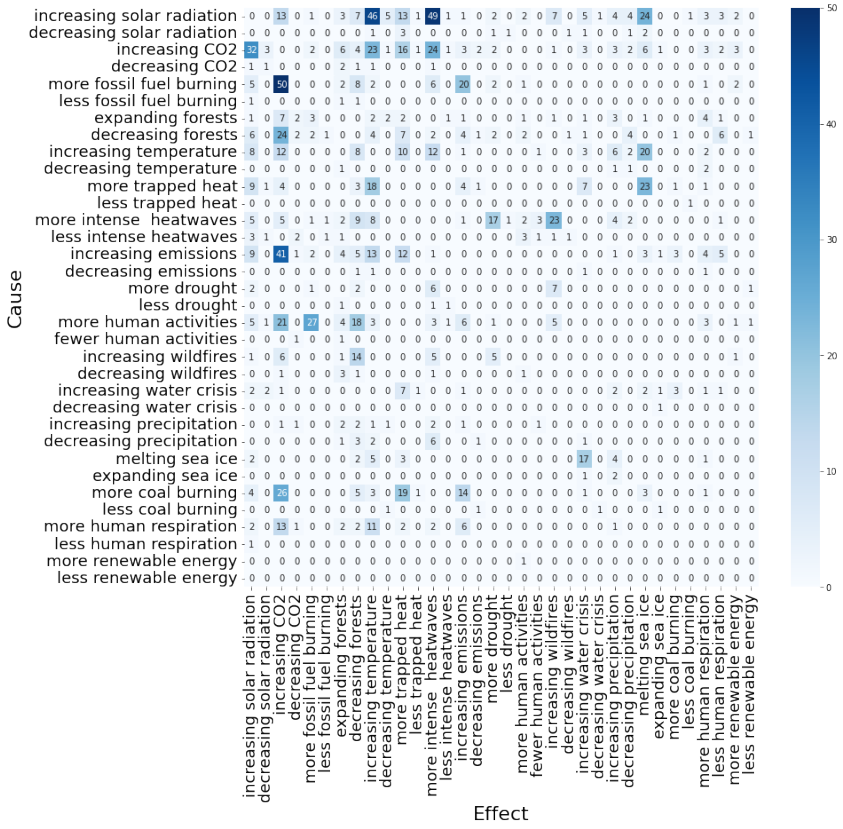

Fig. 4. **The Adjacency Matrix Heatmap Representation of the collected in the formative study.** The cell values represent the total number of votes for that specific causal relation.

7.1 Establishing Credibility Scores

We assigned credibility scores to the causal links based on the methodology proposed in Section 4. Five experts, including two authors of this paper, participated in this stage. The other three experts are climate science experts from our university. They hold PhDs in relevant fields and have been conducting climate science research for at least ten years. Experts independently created their version of the causal networks using the 34 attributes. We provided them with climate change-related literature that is publicly accessible and is more likely to be used by the general population [2]. Figure 3 shows the causal networks created by the experts. The ground truth establishment procedure yielded a dataset of all possible causal links and their credibility scores. Based on our proposed evaluation metrics in Section 4, we conduct an extensive analysis of the causal belief data collected during the experiment. The results are presented next.

[2]https://climate.nasa.gov/
https://www.epa.gov/
https://www.noaa.gov/
https://www.climatecentral.org/

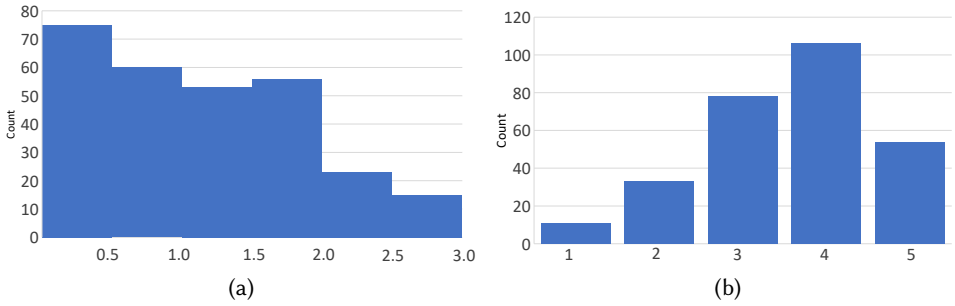

Fig. 5. **The average network credibility scores and the crowd's evaluations/confidence on the causal networks collected in the formative study.** (a) Distribution of Average Network Credibility Scores (0= incorrect link, 3= correct link). (b) Distribution of the crowd's provided confidence scores (1= not confident at all, 5= completely confident).

7.2 Aggregated Evaluation

7.2.1 Combined Network and Total Votes Per Link. We combined all 282 small causal networks by counting the votes for each causal link. Following [82], we present the results as an adjacency matrix heat-map in Figure 4. The top-3 most voted links were, "more fossil fuel burning → increasing CO2", "increasing solar radiation → more intense heatwaves", and "increasing solar radiation → increasing temperature" with 50, 49, and 46 votes, respectively. The combined network is sparse, and most relations have zero votes. The Pearson Correlation Coefficient (r) between the total votes and the credibility scores for the causal links is 0.4 with $p < 10^{-43}$, indicating a moderate consensus between experts and crowd.

7.2.2 Average Network Credibility Scores (ANC) and Average Confidence Score (AC). Figure 5a shows the distribution of ANC scores. The majority of created networks were less credible ($ANC < 2$), suggesting a shallow understanding of climate change topics and occasional worker reluctance. Additionally, Figure 5b displays the distribution of Average Confidence Scores. Less than 15% of cases showed low confidence, while nearly 57% indicated high confidence. These scores reflect the level of certainty associated with the crowd workers' network creations.

The *Aggregated Evaluation* of the data we collected highlights the sparsity of the combined network, the tendency of less credible networks, and the presence of low confidence of the crowd workers in their own created network. These observations motivated us to dive deeper into the data and find the rationales behind them. A closer look at the most popular causal relations revealed that widely acknowledged scientific facts were reflected in the highly voted relations. However, there were also spurious relations, such as "increasing solar radiation → more intense heatwaves" and "increasing solar radiation → increasing temperature". The identified causal relations frequently lacked essential aspects of the climate change issue and resembled arguments used by climate change deniers who attribute it solely to natural factors like the sun [32]. These findings also align with the results of [48], which indicate that people tend to view climate change as an effect of natural variability. It is essential to note that these unfocused relations received lower confidence scores than the more focused ones mentioned above (about 10% lower).

Bogus Cause	More intense heat waves	Less intense heat waves	True Cause	More intense heat waves	Less intense heat waves
Increasing solar radiation	(True)	(False)	More fossil fuel burning	True	False
Decreasing solar radiation	True	False	Less fossil fuel burning	False	True
Bogus Cause	**Increasing temperature**	**Decreasing temperature**	**True Cause**	**Increasing temperature**	**Decreasing temperature**
Increasing solar radiation	(True)	(False)	More fossil fuel burning	True	False
Decreasing solar radiation	True	False	Less fossil fuel burning	False	True

Table 3. **The trial matrices for the two (synonymous) outcome variables heat wave and temperature.** In each matrix the left matrix captures the bogus case and the right is the true cause.

7.3 Formulating the Trial Matrices

For the outcome variable "increasing temperature" (synonymous with "more intense heat waves" in climate science [59]), we identified two competing causes: the simple bogus cause "increasing solar radiation" and the true cause "more fossil fuel burning." The trial matrices for both outcome variables, as shown in Table 3, can be constructed following a similar approach as in Table 2.

The left trial matrix represents the classic bogus cause of "increasing solar radiation." We have added the parentheses to the former since this is a truly imaginary cause as the solar radiation is not really increasing. It indicates that regardless of whether solar radiation increases or not, there will be more intense heat waves. The second column for the opposite outcome, "less intense heat waves," is set to false for both conditions since this outcome is not observed in real life or simulations. The trial matrix on the right represents the true cause of "more fossil fuel burning." It is scientifically established that "more fossil fuel burning" causes "more intense heat waves," while "less fossil fuel burning" eliminates the outcome. The second column, "less intense heat waves," reflects the inverse relationship, as expected in a genuine causal relationship.

It is important to note that our crowd-sourced tool does not provide values for each cell in these trial matrices. Our focus is not on conducting formal experiments with complete matrices, but rather on designing experiments to assess the degree of causal illusion in a general population. We compare the magnitudes of the upper left cells in the trial matrices for the bogus and true causes to gauge the level of causal illusion. Most principled work on causal illusions also largely focuses on these types of results. Additionally, we identify critical links in the causal chain and assess potential knowledge gaps, which are discussed in Section 7.4.2.

7.4 Causal Illusion Detection

The low ANC scores in Section 7.2 indicate a lack of credibility among the crowd-created networks. We explore this further by analyzing the concept of *Causal Illusion*.

7.4.1 Causal Illusions. In Figure 6, we present the *Discrepancy Network* that reveals various levels of discrepancies between the crowd and ground truth networks. The *discrepancy/illusion score* is determined by subtracting the *credibility score* (cs) from the *crowd score* (cr) for each link. A non-zero discrepancy score indicates the presence of a causal illusion. The *crowd score (cr)* is calculated by normalizing the link's total votes, considering the long-tailed distribution of causal links. To

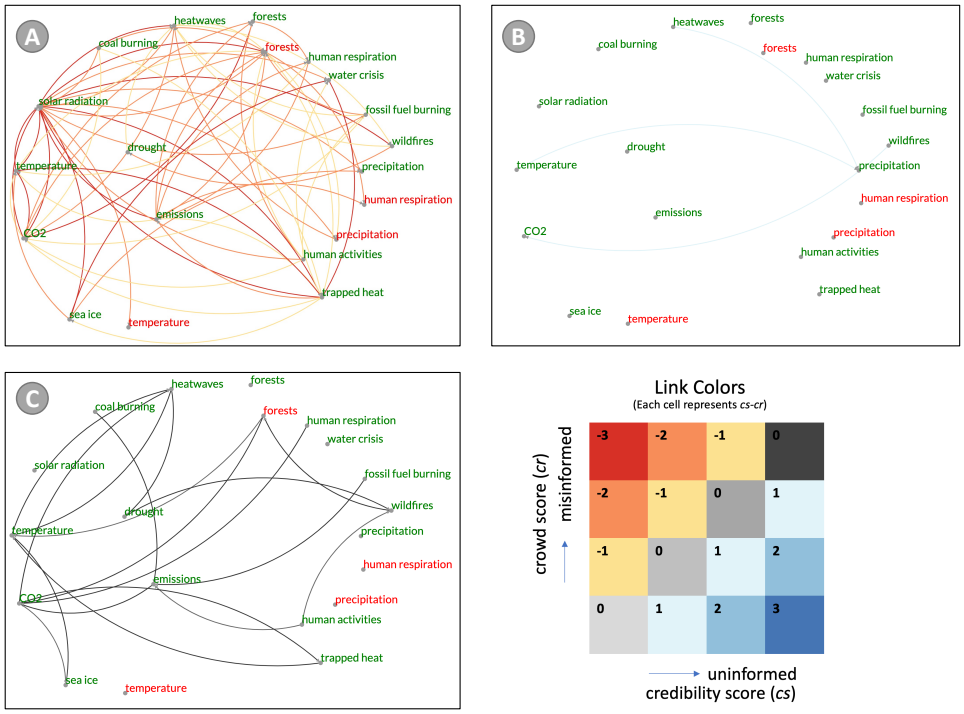

Fig. 6. **The discrepancy networks generated from combined crowd network and ground truth network in the formative study.** Each link color represents the discrepancy between the crowd and ground truth for that specific causal relation. The link colors denote the degree of discrepancy or illusion and the type (*being misinformed* or *being oblivious*). (A) shows the cases of potentially misinformed links, (B) shows the cases of potentially oblivious links, and (C) shows the cases where the crowd correctly predicted the credibility scores. The legend table (row represents crowd score, column represents credibility score) on the right shows the discrepancy/illusion score ($cs - cr$) (in each cell) and the corresponding color. credibility score increases from left to right (→), whereas the crowd score increases from bottom to top (↑). Only the significant links (total vote ≥ 4) and attributes are visible.

assign *crowd scores (cs)* to each link, we utilize a modified equal depth/frequency binning technique. Both *crowd scores (cs)* and *credibility scores (cs)* range from 0 to 3.

Each link color represents either the state of having a causal illusion (being *misinformed* or *oblivious*) or correct (no causal illusion). The colors for these links are coded by the legend on the right. The upper left colors label the state of being misinformed, the lower right colors label the state of being oblivious, and the grey diagonal labels scores where there was agreement. For the latter, we use shading to indicate credibility.

The dominance of *misinformed* links is evident, outnumbering *oblivious* and *correct* links. However, analyzing different discrepancy/illusion scores (Table 4) reveals that correct links constitute almost half of all links (139 out of 281). Notably, when considering significant/visible links, correct links still account for over a quarter (18 out of 61). The lowest grey level shading represents links with minimal votes and is not displayed in the discrepancy network.

Type	Link Color	Discrepancy Score (*cs-cr*)	Count (All)	Count (Visible)
Misinformed	Red	-3	16	16
	Orange	-2	28	28
	Yellow	-1	77	21
Correct	Grey(Darkest)	0 (cs = 3)	14	14
	Grey(Darker)	0 (cs = 2)	4	4
	Grey	0 (cs = 1)	11	0
	Grey(Light)	0 (cs = 0)	110	0
Oblivious	Blue(Very Light)	1	16	3
	Blue(Light)	2	4	0
	Blue	3	1	0
		Total	281	86

Table 4. **Statistics of various discrepancy/illusion scores in the formative study.**

Consistently higher ratios of *misinformed* links emphasize the crowd's vulnerability to misinformation compared to being *oblivious/uninformed*. This vulnerability is particularly evident in the prevalence of misconceptions related to increasing solar radiation among climate-change deniers (Figure 6). Furthermore, other interesting and noteworthy examples, along with more details and analyses, are provided in Appendix A.2. The occurrence of *uninformed* crowd judgments is negligible for significant causal links, underscoring the importance of debunking misinformation and promoting accurate information on complex topics like climate change.

7.4.2 Causal Illusion Quantification. Using the votes obtained for each causal link, we quantified the presence of causal illusion by comparing the bogus cause to the true cause. In Figure 7, we presented the causal links associated with two trial matrices in Table 3, demonstrating the contrasting results. Panel A depicted the simple bogus relation increased solar radiation → increased temperature/more intense heat waves which received a combined vote of 95. In contrast, Panel B showed the more complex true relation starting with increased fossil fuel burning and ends with increased temperature/more intense heat waves. Various traversals were observed in which the wisdom of the crowd navigated this chain. The most accurate 4-hop path (increased fossil fuel burning → increasing emissions → increasing CO2 → more trapped heat → increased temperature/more intense heat waves) received a support of 16 votes, assuming the weakest link as the defining one, or 23.75 votes on average. Additionally, 3-hop and 2-hop paths also emerged, each offering varying degrees of accuracy. Evaluating the degree of illusion, we found that the ratio of votes for the bogus cause to the true cause ranged from 95/23.75=4 to 95/14=5.94, indicating a stronger belief in the causal illusion. On the other hand, if we compare the dominant true causal knowledge with the bogus cause we get a ratio of about 2.

7.5 Design Issues

7.5.1 Selection Bias. One important issue in the initial interface was the potential for selection bias due to the pre-determined order and the selective and gradual appearance of attributes in the interface. The order in which attributes are presented can influence participants' perceptions and choices. Further, it is possible that the positioning of certain attributes based on Word2Vec coordinates influenced participants to perceive and associate them more favorably compared to others. This may have contributed to the sparse nature of the combined network (Figure 4).

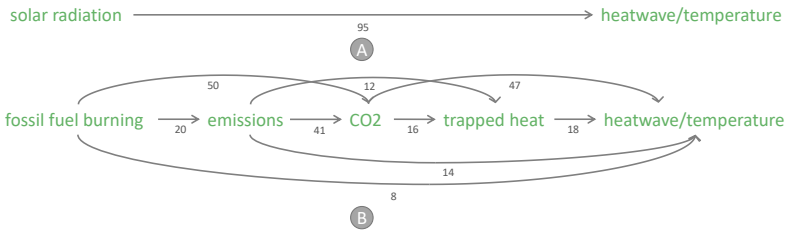

Fig. 7. **Extracted causal chains in the formative study for the bogus cause and the true cause.** Panel A is the simple link that connects the bogus cause with the outcome. This is the link selected by the fraction of the crowd who fell victim to the causal illusion. Panel B are several pathways of varying complexity by which the crowd has linked the true cause with the outcome. The numbers below each link are the number of votes the link received in the compound causal network. The green color indicates rising values.

7.5.2 Longer Completion Time than Anticipated. The average completion time was 30 minutes, exceeding the estimated duration of 15 minutes, which directly contradicted our design goal, **DG3**. One reason behind that could be the three rounds required to complete the task. Participants had to perform a total of seven clicks in various parts of the interface to create a single link. The visual representation of node-link diagrams could also contribute to the longer completion times.

8 RE-DESIGNED AND FINAL INTERFACE

The formative study produced encouraging results, as the crowd demonstrated the ability to create meaningful networks while exhibiting evidence of various types of causal illusions. However, we also identified two major design issues. Based on these findings, we revised the design of the protocol and interface for collecting causal networks from crowd workers. We had three specific goals in mind. Firstly, we aimed to eliminate any potential selection bias that may have influenced participant responses. Secondly, we sought to streamline the data collection process by reducing the number of micro-tasks required and the amount of completion time. Lastly, we decided to enforce a structured process for recruiting participants and performing post-hoc quality control.

8.1 Collection Interface Modules

The redesigned interface utilized a similar web-based technology as our initial interface. We excluded the attribute "renewable energy" from our input dataset, as it had a low number of votes in the formative study. The interface also has modules similar to the initial interface. We describe major changes to the modules below.

8.1.1 Demographics Survey Module. We decided to replace the previous questions gauging participants' self-assessed knowledge and awareness with a more standardized question set since crowdworkers may over or underestimate their knowledge about climate change [42]. We decided to use the Six Americas Super Short Survey (SASSY) [18]. These questions classify individuals into six categories representing different levels of climate change awareness: Alarmed, Concerned, Cautious, Disengaged, Doubtful, and Dismissive.

Note that SASSY captures people's attitudes towards climate change, not their knowledge level. Our target audience is the *general population*, thus neither knowledge nor attitude is an influential factor in the study. Nevertheless, we wanted to recruit people with diverse opinions and beliefs about climate change to validate our method and find a wide range of causal beliefs and illusions. With the lack of a standardized method for measuring knowledge about climate change, we believe SASSY is a reasonable proxy to determine the diversity of our participants' pool.

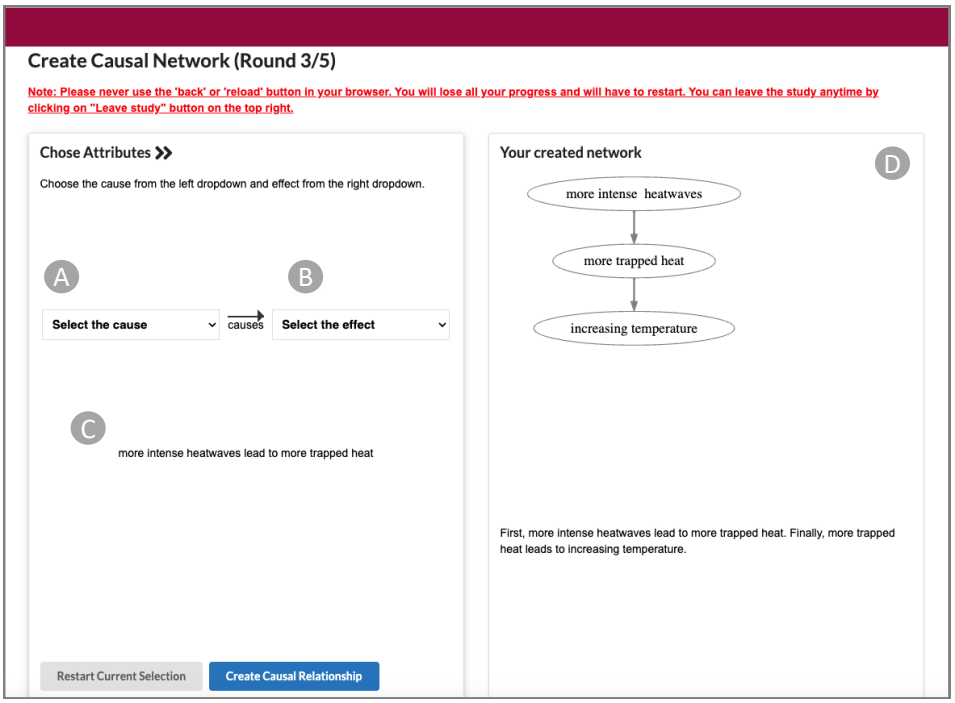

Fig. 8. **Overview of the redesigned collection interface**. This example shows a causal network is being created. A shows the "cause" drop-down and B shows the "effect" drop-down. C shows the most recently created causal link as text. D contains the emerging causal network along with its textual narrative.

8.1.2 Causal Network Creation Module. We made four major changes to the Causal Network Creation Module (see Figure 8). *First*, instead of presenting the attributes using a node-link diagram, we provide two drop downs for users to select causes and effects (Figure 8A and B). This design reduces the possibilities of biases induced by the position of the nodes in a node-link diagram and scales the interface to any number of attributes. *Secondly*, instead of randomly selecting a subset of attributes, we presented all attributes to the users in the drop-downs. This design choice again reduces any possibilities of selection bias. *Thirdly*, since a user can see the whole set of attributes from the beginning, we only asked participants to complete a single session (creating five relations), eliminating the need for the additional two sessions. This design choice significantly reduces the completion time. *Finally*, even though users do not create networks using node-link diagrams, we show the created diagrams to the users using GraphViz [26], a well-known graph visualization library that has built-in rendering mechanisms (Figure 8D). We also reduced the number of micro-tasks for optimizing the completion time. For example, trend selection was a separate micro-task in the initial interface, but now the trends are embedded with the attribute names. Users do not need to select trends separately, reducing two micro-tasks in total (one for selecting cause and one for effect). We attached a video to demonstrate the overall workflow. Appendix B.1 describes the workflow in more detail.

9 FINAL STUDY

We recruited 72 crowd workers from Amazon Mechanical Turk (AMT) and 60 workers on Prolific (132 workers in total) to collect causal perceptions on climate change using our final interface. We

excluded the work of 31 workers due to incomplete and fraudulent results, totaling to 101 valid responses. Successful crowd workers were compensated $3.75 on AMT and $3 on Prolific upon completion. The compensation amount on Prolific was determined automatically based on the estimated completion time, while on AMT we estimated it ourselves. The average completion time was around 12 minutes. The study design and protocol remained similar to our formative study, except for the following changes. We mention the detailed protocol in Appendix B.1.

9.1 Crowd Workers' Validation Process and Post-hoc Quality Control

In the final study, in contrast to the formative study, we implemented a rigorous quality control process to ensure data integrity. First, we noticed from the formative study that while we designed the *crediblilty score* to measure causal beliefs and illusions, it can also be used to filter out potential fraudulent or random causal relations. For example, an excessive number (3 or more out of 5) of non-credible links (credibility score = 0) could indicate randomly created relations or fraudulent behaviors. We flagged such relations and crowd workers in the final study. Then, the research team manually reviewed each case. During the review, we examined patterns of inconsistencies, repeated responses, and indications of random or careless selection. Based on the manual review, we identified and excluded workers who produced unreliable or fraudulent data. We provide two examples of such data in Figure 9-A and B. Note that the lack of credibility (a score of 0) does not necessarily mean fraudulent behavior. We accepted submission as long as the overall network exhibited a pattern of understanding and relevance, even if some links had a credibility score of 0 (Figure 9-C). We interpreted these results as beliefs stemming from potentially flawed understanding, which are of potential interest to us, and accepted the results. Thus, the combined networks had links with credibility scores spanning from 0 to 3.

We conducted the study in phases, 10 participants at a time, allowing us to calculate node exploration in the aggregated network and perform quality control at each phase. All 31 workers whose results were rejected due to incomplete or fraudulent data were from AMT. Thus, we decided to conduct the rest of the study on Prolific, as it is known for its more stringent participant vetting process and higher-quality data. We ensured all participants finished the study exactly once. Therefore, every participant from AMT and Prolific constructed their own network from the ground up without building upon one another's work. This led us to have 101 valid workers. We present various aspects of the crowd workers' demographics in Figure 10. Geographically, we only collected results from the United States.

9.2 Crowd Workers' View on Climate Change

We employed the "*Six Americas Super Short SurveY (SASSY)*" Group Scoring Tool to segment our participants into different groups based on their responses to climate change [18]. We present the results in Figure 11. The majority of workers (36) were "Alarmed" and expressed high concern about climate change. There were also significant numbers of workers in the "Dismissive" group (25) who held dismissive or denying attitudes. Other categories included "Concerned" (11), "Cautious" (10), "Doubtful" (16), and "Disengaged" (3). The group scoring tool also allowed us to compare the groups to national averages (Figure 11). Compared to national averages from December 2022, we observed some variations in the distribution of attitudes toward climate change. Nevertheless, all groups were present in our participant pool, reflecting a range of perspectives and attitudes toward climate change.

9.3 Stopping Criteria

We conducted the study in phases (10 participants at a time). After each phase, we analyzed the combined network created by the crowd and compared it with the previous phases. We also checked

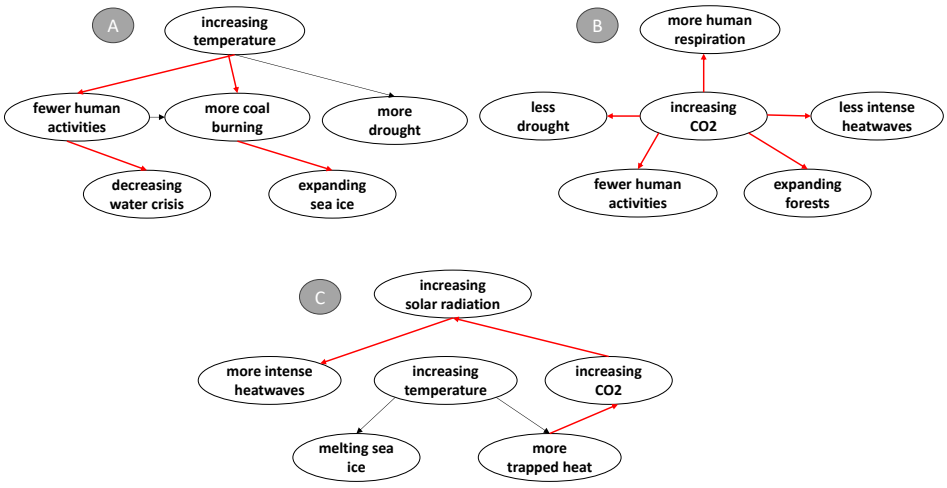

Fig. 9. **Three examples of flagged data collected during the final study**. **A** and **B** are fraudulent data that were rejected, whereas **C** was accepted. Network **A** has 4 and Network **B** has 5 links that show patterns of inconsistencies and indications of random or careless selection (marked in red). For example, in Network **A**, "increasing temperature → more coal burning" is a spurious link, the scientific evidence supports the opposite relation. In Network **B**, "increasing CO2 → less intense heatwaves" is another example of such a spurious link, as increasing CO2 levels contribute to intensified heatwaves, not diminished ones. Links inside one network also mostly do not bear any relevance to each other in Networks **A** and **B**. On the other hand, Network **C**, has three scientifically invalid links (marked in red), but all links are fitting and express a coherent narrative.

Fig. 10. **Demographics of the crowd workers in the final study.** Y-axes represent counts for each category.

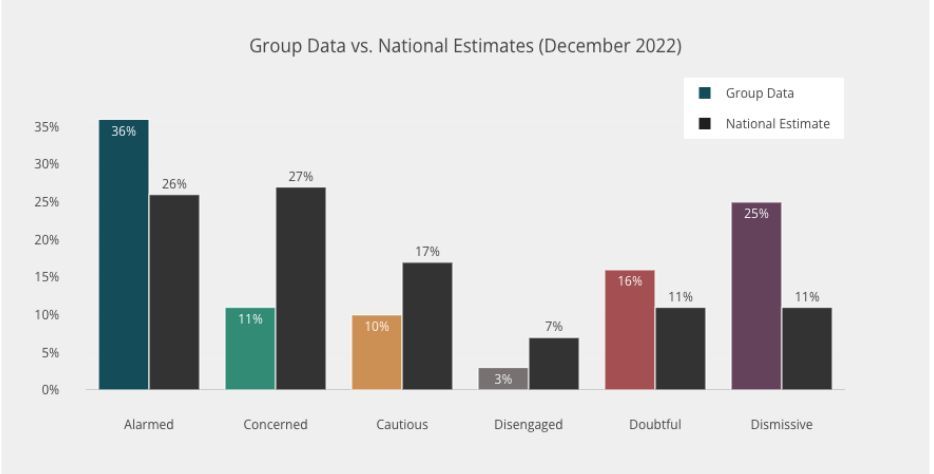

Fig. 11. **Crowd workers' view on climate change, compared to the national estimates.** The black bars represent the national estimates and the colored bars represent our sample. The chart was generated using the group scoring tool provided by [18].

how many nodes/attributes have been explored in the combined network. We stopped the study once the network was saturated, displaying minimal changes from previous phases. We present the final node exploration status in Appendix B.2. Additionally, we considered the representation of the participants in terms of the SASSY groups. We wanted representatives from each group.

10 FINDINGS FROM THE FINAL STUDY

The analysis pipeline in the final study remained consistent with that of the formative study. We present the findings below.

10.1 Aggregated Evaluation

10.1.1 Combined Network and Total Votes Per Link. We present the adjacency matrix of the 101 combined networks in Figure 12. The top-3 most voted links were, "increasing temperature → melting sea ice", "increasing CO2 → more trapped heat," and "more trapped heat → increasing temperature" with 24, 21, and 18 votes, respectively. The combined network is sparse, and most relations have zero votes, similar to the formative study. The Pearson Correlation Coefficient (r) between the total votes and the credibility scores for the causal links is 0.63 with $p < 10^{-105}$. This indicates a stronger alignment between the crowd and expert consensus compared to the formative study. Although the level of consensus remains moderate, the stronger correlation suggests that the crowd was able to create networks that contain more scientifically accurate causal links in general.

10.1.2 Average Network Credibility Scores (ANC) and Average Confidence Scores (AC). Figure 13a shows the distribution of ANC scores, which reflects that the majority of the crowd workers created less credible networks ($ANC < 2$) in general. We present the confidence scores in Figure 13b, which reflects that in less than 18% of the cases, the workers did not feel very confident (Confidence Score 1-2) in their own created network and nearly 54% of the time, they felt positively confident (Confidence Score 4-5).

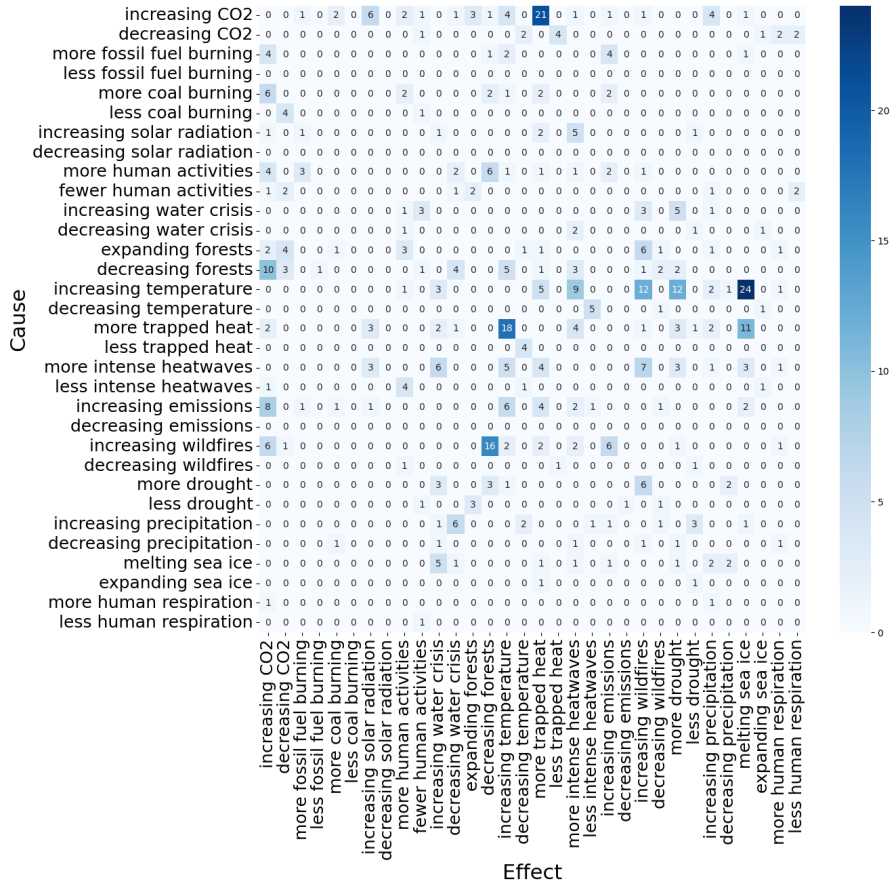

Fig. 12. **The Adjacency Matrix Heatmap Representation of the 101 causal networks collected in the final study.** The cell values represent the total number of votes for that specific causal relation.

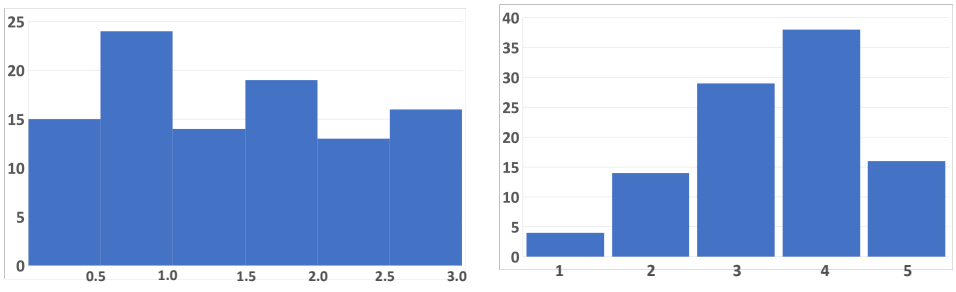

Fig. 13. **The actual average network credibility scores and the crowd's evaluations/confidence on the causal networks collected in the final study.** (a) Distribution of Average Network Credibility Scores (0= incorrect link, 3= correct link). (b) Distribution of the crowd's provided confidence scores (1= not confident at all, 5= completely confident).

10.2 A Closer Look into the Most Popular Causal Relations

Similar to the formative study, we mention the most noteworthy aspect of the most popular causal links below.

- Similar to the formative study, the most voted causal relations in the final study reflected widely acknowledged scientific facts. The top relation (24 votes, average confidence (AC) = 3.5, credibility score (CS) = 3) was "increasing temperature → melting sea ice". Other top relations attributed the "more trapped heat" to "increasing CO2" (21 votes, 3.5 AC, 3 CS) and "increasing temperature" to "more trapped heat (18 votes, 3.3 AC, 3 CS).
- There was evidence of recognizing underlying factors of climate change, such as linking "more human activities" to "decreasing forests" (6 votes, 3.7 AC, 2 CS) and "increasing CO2" (4 votes, 3.5 AC, 2 CS) and considering "increasing CO2" as an effect of "decreasing forests" (10 votes, 3.7 AC, 3 CS) and "increasing emissions" (8 votes, 3.5 AC, 3.5 CS) [20, 28].
- There was also evidence of physical understanding; many marked that "increasing wildfire" leads to "decreasing forests" (16 votes, 3.5 AC, 3 CS) and "increasing emissions" (6 votes, 3.7 AC, 3 CS) and "increasing CO2" (6 votes, 3.5 AC, 2 CS), and, in turn, "increasing wildfires" is caused by "increasing temperature" (12 votes, 3.2 AC, 2 CS) and "more intense heatwaves" (7 votes, 4.14, 2 CS).
- Among the causal links that received at least 10 votes, all were highly credible links (CS=3) except for 2 links: "increasing temperature" → "increasing wildfires" and "more trapped heat" → "melting sea ice" (CS=2). Interestingly, these links also received the lowest AC scores (3.2 and 2.8, respectively) compared to others, exhibiting a parallel trend to the formative study findings. While these links are partially correct, they lack crucial mediators. It is important to acknowledge that wildfires are primarily caused by dry weather resulting from *drought*, which is a consequence of rising temperatures. Similarly, the consistent trapping of heat contributes to *increasing temperatures*, which leads to the melting of sea ice [25, 49].

10.3 Causal Illusion Detection

In Figure 14 we present the *Discrepancy Network* along with various levels of discrepancies. Similar to the formative study, the ratio of *misinformed* links consistently stays higher (47 out of 92 visible links). In contrast to the formative study, the results of the current study reveal a multitude of cases where the crowd demonstrated a significant degree of obliviousness, with 26 out of 92 visible links exhibiting this phenomenon. We present the statistics of various values of different discrepancy/illusion scores in Table 5.

In Figure 15, we present some noticeable cases within the discrepancy network. The crowd selected increasing solar radiation as a cause for more intense heatwaves, this misconception is consistent among the participants of both studies (Figure 15-A). The crowd also seemed to think that expanding forest causes increasing wildfires (Figure 15-B), whereas, the relationship between expanding forests and increased wildfires is more nuanced. Factors such as climate conditions, human activities, and forest management practices play significant roles in determining wildfire risk [25, 58].

Several intriguing instances of obliviousness were observed within the crowd as depicted in Figure 15-X, Y, Z. Notably, the crowd exhibited a preference for attributing emissions to increasing wildfires, which is indeed a valid relationship. However, there was a comparative disregard for the influence of fossil fuel burning and coal burning on emissions. Additionally, a certain level of unawareness was evident regarding the interplay between variables such as drought, precipitation, and temperature-related factors like trapped heat and heatwaves. This highlights the need to

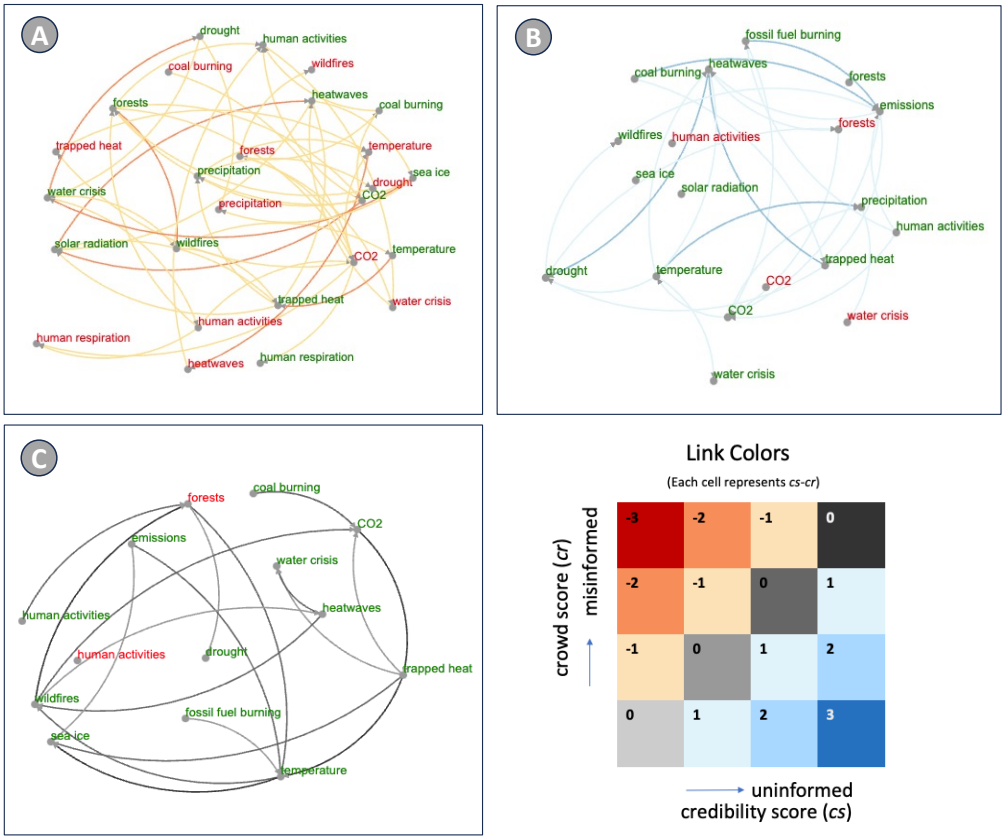

Fig. 14. **The discrepancy networks generated from the combined crowd network and the ground truth network in the final study.** Each link color represents the discrepancy between the crowd and ground truth for that specific causal relation. The link colors denote the degree of discrepancy or illusion and the type (*being misinformed* or *being oblivious*). (A) shows the cases of potentially misinformed links, (B) shows the cases of potentially oblivious links, and (C) shows the cases where the crowd correctly predicted the credibility scores.

prioritize the dissemination of accurate information concerning the impact of rising temperatures on drought conditions.

10.3.1 Causal Illusion Quantification. Figure 16 illustrates the causal links associated with the two trial matrices described in Section 7.3 (Table 3). The strongest 4-hop path (more fossil fuel burning → increasing emissions → increasing CO2 → more trapped heat → increased temperature/more intense heat waves) received 4 votes (based on the weakest link criterion) or 13.75 votes (based on the average link criterion). The optimal 3-hop path (more fossil fuel burning → increasing CO2 → more trapped heat → increased temperature/more intense heat waves) garnered 4 votes (weakest link criterion) or 15.67 votes (average link criterion). The best 2-hop path (more fossil fuel burning → increasing CO2 → increased temperature/more intense heat waves) obtained 4 votes (weakest link criterion) or 4.5 votes (average link criterion). Lastly, the simplest 1-hop path (more fossil fuel burning → increasing temperature/more intense heat waves) received 2 votes. To assess the degree

Type	Link Color	Discrepancy Score (*cs-cr*)	Count (All)	Count (Visible)
Misinformed	Red	-3	0	0
	Orange	-2	7	7
	Yellow	-1	40	40
Correct	Grey(Darkest)	0 (cs = 3)	4	4
	Grey(Darker)	0 (cs = 2)	9	9
	Grey	0 (cs = 1)	6	6
	Grey(Light)	0 (cs = 0)	75	0
Oblivious	Blue(Very Light)	1	27	21
	Blue(Light)	2	9	5
	Blue	3	4	0
		Total	181	92

Table 5. **Statistics of various discrepancy/illusion scores in the final study.**

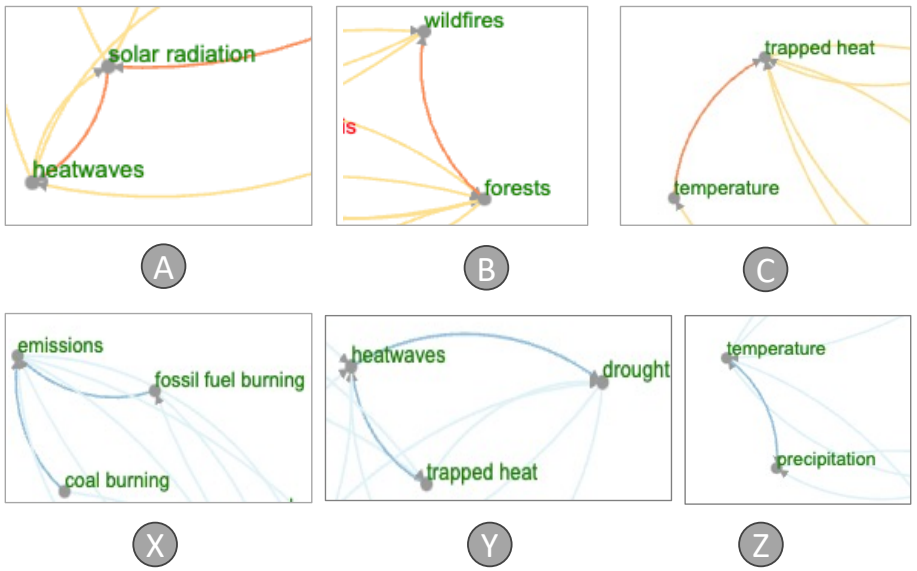

Fig. 15. ***Misinformed* and *Oblivious* cases in discrepancy network.**

of illusion between the bogus cause and the true cause, we can examine the vote ratios. For the most accurate 4-hop path, this ratio is 5/4 = 1.25 for the weakest link criterion and 5/13.75 = 0.36 for the average criterion. Although the weakest link criterion suggests a mild degree of illusion, the number of votes is too low. Another encouraging finding is that a significant number of workers correctly identified the direct cause of increased temperature/more intense heat waves, which is more trapped heat (21 votes).

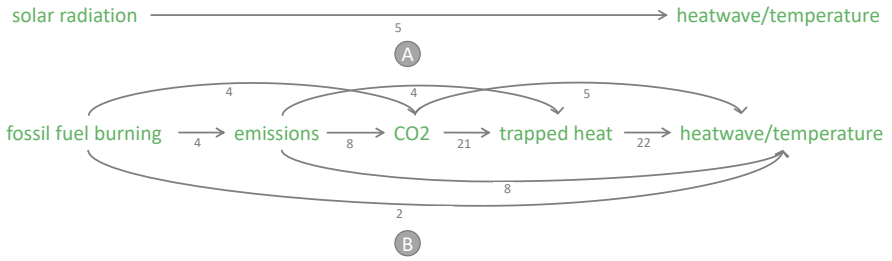

Fig. 16. **Replication of Figure 7 in the final study.**

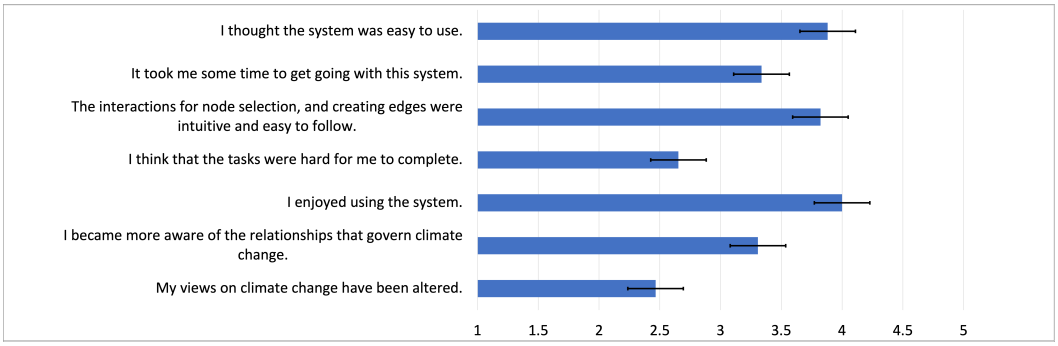

Fig. 17. **Average ratings provided by the crowd workers of the final study on 7 subjective usability and knowledge-related statements.** Error bars represent standard errors. (1=Strongly Disagree, 5=Strongly Agree).

10.4 Interface Usability and Knowledge

Figure 17 presents crowd workers' agreement on the usability and learning statements. The results reflect an overall positive outlook toward the usability of our system. For the knowledge-related statements, the majority of workers felt neutral to positive towards the statement "I became more aware of the relationships that govern climate change" and disagreed with the statement "My views on climate change have been altered". We think this happened because they might not have felt comfortable agreeing to such a strong statement.

11 DISCUSSION, LIMITATIONS, AND FUTURE WORK

We briefly summarize the study findings, design implications, and limitations of our work below.

11.1 Reflecting on the User Studies

The two studies show that Belief Miner is data-driven and agnostic to the method used to collect the data. Regardless of the collection mechanism, Belief Miner can extract causal illusions given a set of causal networks collected from the crowd and ground truth collected from experts. Thus, we were able to apply the same method in both studies, even though the underlying protocol and interfaces were different.

We found a wide range of causal beliefs and illusions in the studies. Some findings from the studies align while others do not. For example, in both studies, participants with causal illusions assigned lower confidence scores to their networks. However, we could not replicate some results

from the formative study in the final study. For example, we did not find the causal illusion reported in Figure 7 from the formative study in the final study. Several factors could contribute to this phenomenon: the redesigned interface, the participants' pool, or the changes in the general knowledge about climate change within the timeframe of the studies (Fall 2021 to Spring 2023).

11.2 Design Implications

11.2.1 Belief Miner as an Intervention against Causal Illusion. In our current system, we collect the causal beliefs from crowd workers through the interactive visual interface and detect the causal illusions later. Both of our studies show the phenomenon of people rating their own causal networks containing spurious causal links with low confidence scores once they see the complete picture. We see this as an indication of people correcting or educating themselves while they see the externalized version of their mental causal model. Motivated by this, we envision using Belief Miner as an intervention tool against causal illusion. Several crowdsourcing methods could be useful here. For example, seeing other people's beliefs often positively affect a crowdsourcing task [43, 82]. In our case, we can expose the controversy or disagreement about a causal relation at the time of the experiment by looking at the networks already created by others. Another potential solution is enabling peer review [77], allowing crowdworkers to provide feedback to each other. Other potential solutions include automatically extracting digestible scientific documents related to the relevant causal attributes. We believe these additions will facilitate informed decision-making and will promote scientific thinking which is identified as the best defense against causal illusions [54].

11.2.2 From Causal Illusion to Misconception. Belief Miner investigates the concept of causal illusion, which is related to people's inherent bias to draw connections between coincidental events. A closely related concept is misconception, which is the inaccurate or wrong interpretation of concepts [69]. The terms misconception and misinformation are often used interchangeably. While misconception generally comes from a lack of knowledge, misinformation is often deliberately created for deception and spread intentionally or unintentionally [78]. Existing misconception and misinformation discovery methods mainly rely on natural language processing (NLP) and machine learning (ML). These methods fall under the broader category of content-based detection [4, 55], context-based detection [46], propagation-based detection [38], etc. We believe our method could provide a realistic tool to measure misconceptions where content-based analysis is not feasible.

11.2.3 What Kinds of Causal Illusions Appear Together and Who Falls Victim to Them? Another interesting future direction is utilizing the demographic profile of the crowdworkers and the cases of potential causal illusions to identify groups of illusions along with the population groups and their geo-locations who fall victim to them. One way of doing this is to find all possible cases and cluster them based on their structures, i.e., the attributes, type of the link, and level of causal illusions (introduced in Section 7.4). Another potential way of finding such groups can be using frequent itemset mining algorithms such as Apriori [2]. In the case of the causal belief dataset, each causal network made by a particular worker consisting of a collection of causal relations and their casual illusion levels represents one transaction. These clusters/frequent itemsets can potentially bring out population groups susceptible to that specific groups of causal illusions. Eventually, these population groups can represent different schools of thought. Therefore, we think it is worth exploring this idea as an extension of our current methodology.

11.2.4 Belief Miner for Causal ML. Prior research has used causal crowdsourcing as a way to develop training datasets for machine learning [82]. While our work focuses on behavioral analysis, we believe detecting causal illusions would be useful for machine learning too. If not detected, causal illusions can lead to incorrect decision-making and machine-intelligence.

12 CONCLUSION

We presented *Belief Miner*, a methodology for collecting and evaluating the crowd's causal beliefs and discovering causal illusions. We developed two interactive interfaces to collect such beliefs, where people can create small causal networks to create a large causal network collectively. Two separate crowdsourcing studies show that all participants successfully created the small networks using our interactive interface, except for a few dropouts. Our evaluation methodology can find potential discrepancies and illusions in the causal relations created by the crowd. We hope our work will start the discussion on different methods of analyzing people's causal opinions and beliefs and what they may reveal about the people themselves.

ACKNOWLEDGMENTS

This research was partially supported by NSF grant IIS 1941613 and IIS 1527200.

REFERENCES

[1] Jon Agley, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo. 2022. Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior research methods* 54, 2 (2022), 885–897.

[2] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Santiago, Chile, 487–499.

[3] Lorraine G Allan. 1980. A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society* 15, 3 (1980), 147–149.

[4] Adrien Benamira, Benjamin Devillers, Etienne Lesot, Ayush K. Ray, Manal Saadi, and Fragkiskos D. Malliaros. 2019. Semi-Supervised Learning and Graph Neural Networks for Fake News Detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Vancouver, British Columbia, Canada) *(ASONAM '19)*. Association for Computing Machinery, New York, NY, USA, 568–569. https://doi.org/10.1145/3341161.3342958

[5] Daniel Berenberg and James P. Bagrow. 2018. Efficient Crowd Exploration of Large Networks: The Case of Causal Attribution. *Proc. ACM Hum. Comput. Interact.* 2, CSCW (2018), 24:1–24:25. https://doi.org/10.1145/3274293

[6] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, October 3-6, 2010*. ACM, New York, NY, USA, 313–322. https://doi.org/10.1145/1866029.1866078

[7] Jeffrey P Bigham, Michael S Bernstein, and Eytan Adar. 2015. Human-computer interaction and collective intelligence. *Handbook of collective intelligence* 57 (2015).

[8] Fernando Blanco, Helena Matute, and Miguel A. Vadillo. 2012. Mediating role of activity level in the depressive realism effect. (2012).

[9] Fernando Blanco, Helena Matute, and Miguel A Vadillo. 2013. Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior* 41, 4 (2013), 333–340.

[10] Enrico Blanzieri. 2012. The role of causal beliefs in technology-supported policy. *IFAC Proceedings Volumes* 45, 10 (2012), 171–176.

[11] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[12] Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing*. ACL, 286–295. https://aclanthology.org/D09-1030/

[13] MM Cartwright. 2011. Alternative medicine & the death of Steve Jobs. *Psychology Today. October* 21 (2011).

[14] Tommaso Caselli and Oana Inel. 2018. Crowdsourcing StoryLines: Harnessing the Crowd for Causal Relation Annotation. In *Proceedings of the Workshop Events and Stories in the News 2018*. Association for Computational Linguistics, Santa Fe, New Mexico, U.S.A, 44–54. https://aclanthology.org/W18-4306

[15] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*. ACM, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[16] Chia-En Chiang, Yu-Chun Chen, Fang-Yu Lin, Felicia Feng, Hao-An Wu, Hao-Ping Lee, Chang-Hsuan Yang, and Yung-Ju Chang. 2021. "I Got Some Free Time": Investigating Task-Execution and Task-Effort Metrics in Mobile

Crowdsourcing Tasks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 648, 14 pages. https://doi.org/10.1145/3411764.3445477

[17] Arjun Choudhry, Mandar Sharma, Pramod Chundury, Thomas Kapler, Derek W. S. Gray, Naren Ramakrishnan, and Niklas Elmqvist. 2021. Once Upon A Time In Visualization: Understanding the Use of Textual Narratives for Causality. *IEEE Transactions on Visualization and Computer Graphics* 27 (2021), 1332–1342.

[18] Breanne Chryst, Jennifer Marlon, Xinran Wang, Sander van der Linden, Edward Maibach, Connie Roser-Renouf, and Anthony Leiserowitz. [n. d.]. Six Americas Super Short Survey (SASSY!). https://climatecommunication.yale.edu/visualizations-data/sassy/. (Accessed on 07/10/2023).

[19] John Joon Young Chung, Jean Y. Song, Sindhu Kutty, Sungsoo (Ray) Hong, Juho Kim, and Walter S. Lasecki. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 62:1–62:25. https://doi.org/10.1145/3359164

[20] Climate.gov. [n. d.]. Climate Change: Atmospheric Carbon Dioxide. https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide. (Accessed on 07/10/2022).

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[22] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[23] Steven Dow, Anand Pramod Kulkarni, Scott R. Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012*. ACM, New York, NY, USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[24] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*. ACM, New York, NY, USA, 2399–2402. https://doi.org/10.1145/1753326.1753688

[25] Drought.gov. 2022. Wildfire Management. https://www.drought.gov/sectors/wildfire-management (Accessed: 2023-07-10).

[26] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. 2002. Graphviz—open source graph drawing tools. In *Graph Drawing: 9th International Symposium, GD 2001 Vienna, Austria, September 23–26, 2001 Revised Papers 9*. Springer, 483–484.

[27] Whitney Fleming, Adam L. Hayes, Katherine M. Crosman, and Ann Bostrom. 2021. Indiscriminate, Irrelevant, and Sometimes Wrong: Causal Misconceptions about Climate Change. *Risk Analysis* 41, 1 (2021), 157–178. https://doi.org/10.1111/risa.13587 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.13587

[28] American Forests. 2022. Forests as Carbon Sinks. https://www.americanforests.org/article/forests-as-carbon-sinks/ (Accessed: 2022-02-04).

[29] Ian Freckelton. 2012. Death by homeopathy: issues for civil, criminal and coronial law and for health service policy. *Journal of Law and medicine* 19, 3 (2012), 454–478.

[30] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1631–1640. https://doi.org/10.1145/2702123.2702443

[31] Bhavya Ghai and Klaus Mueller. 2023. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 473–482. https://doi.org/10.1109/TVCG.2022.3209484

[32] NASA Global Climate Change. 2021. Is the Sun causing global warming? https://climate.nasa.gov/faq/14/is-the-sun-causing-global-warming/ Accessed: 2022-06-04.

[33] National Health and Medical Research Council. 2015. *NHMRC Information Paper: Evidence on the effectiveness of homeopathy for treating health conditions*. National Health and Medical Research Council.

[34] Md. Naimul Hoque and Klaus Mueller. 2022. Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making. *IEEE Trans. Vis. Comput. Graph.* 28, 12 (2022), 4728–4740. https://doi.org/10.1109/TVCG.2021.3102051

[35] John Joseph Horton and Lydia B Chilton. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*. 209–218.

[36] Panagiotis G. Ipeirotis, Foster J. Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '10, Washington DC, USA, July 25, 2010*. ACM, New York, NY, USA, 64–67. https://doi.org/10.1145/1837885.1837906

[37] Sanjay Kairam and Jeffrey Heer. 2016. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social*

Computing, CSCW 2016, San Francisco, CA, USA, February 27 - March 2, 2016. ACM, New York, NY, USA, 1635–1646. https://doi.org/10.1145/2818048.2820016

[38] Jooyeon Kim, Dongkwan Kim, and Alice Oh. 2019. Homogeneity-Based Transmissive Process to Model True and False News in Social Networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 348–356. https://doi.org/10.1145/3289600.3291009

[39] Yongsung Kim, Darren Gergle, and Haoqi Zhang. 2018. Hit-or-Wait: Coordinating Opportunistic Low-Effort Contributions to Achieve Global Outcomes in On-the-Go Crowdsourcing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173670

[40] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence, Italy, April 5-10, 2008*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[41] Aniket Kittur, Susheel Khamkar, Paul André, and Robert E. Kraut. 2012. CrowdWeaver: visually managing complex crowd work. In *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012*. ACM, New York, NY, USA, 1033–1036. https://doi.org/10.1145/2145204.2145357

[42] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[43] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/2047196.2047202

[44] Ranjay A. Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A. Shamma, Li Fei-Fei, and Michael S. Bernstein. 2016. Embracing Error to Enable Rapid Crowdsourcing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3167–3179. https://doi.org/10.1145/2858036.2858115

[45] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 4066–4076. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[46] Sejeong Kwon and Meeyoung Cha. 2014. Modeling Bursty Temporal Pattern of Rumors. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 650–651. https://ojs.aaai.org/index.php/ICWSM/article/view/14494

[47] Walter S. Lasecki, Jeffrey M. Rzeszotarski, Adam Marcus, and Jeffrey P. Bigham. 2015. The Effects of Sequence and Delay on Crowd Work. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1375–1378. https://doi.org/10.1145/2702123.2702594

[48] Anthony Leiserowitz, Edward W Maibach, Seth Rosenthal, John Kotcher, Parrish Bergquist, Matthew Ballew, Matthew Goldberg, and Abel Gustafson. 2019. Climate change in the American mind: April 2019. *Yale University and George Mason University. New Haven, CT: Yale Program on Climate Change Communication* (2019).

[49] World Wild Life. [n. d.]. Why are glaciers and sea ice melting? https://www.worldwildlife.org/pages/why-are-glaciers-and-sea-ice-melting. (Accessed on 07/10/2023).

[50] Ching Liu, Juho Kim, and Hao-Chuan Wang. 2018. ConceptScape: Collaborative Concept Mapping for Video Learning. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173961

[51] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[52] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. 2007. Mediation analysis. *Annual review of psychology* 58 (2007), 593.

[53] Narges Mahyar, Michael R. James, Michelle M. Ng, Reginald A. Wu, and Steven P. Dow. 2018. CommunityCrit: Inviting the Public to Improve and Evaluate Urban Design Ideas through Micro-Activities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173769

[54] Helena Matute, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A Vadillo, and Itxaso Barberia. 2015. Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in psychology* 6 (2015), 888.

[55] Md Abdullah Al Mazid and Zaima Zarnaz. 2022. Climate Change Myths Detection Using Dynamically Weighted Ensemble Based Stance Classifier. In *Proceedings of the 2nd International Conference on Computing Advancements* (Dhaka, Bangladesh) *(ICCA '22)*. Association for Computing Machinery, New York, NY, USA, 277–283. https://doi.org/10.1145/3542954.3542995

[56] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[57] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/2047196.2047198

[58] Nps.gov. 2022. Wildfire Causes and Evaluations. https://www.nps.gov/articles/wildfire-causes-and-evaluation. (Accessed: 2023-07-10).

[59] World Meteorological Organization. [n. d.]. Heatwaves. https://www.britannica.com/science/heat-wave-meteorology. (Accessed on 07/10/2022).

[60] Trishan Panch, Heather Mattie, and Rifat Atun. 2019. Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health* 9, 2 (Dec. 2019), 010318.

[61] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect.* Basic books.

[62] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[63] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376403

[64] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 83 (apr 2022), 22 pages. https://doi.org/10.1145/3512930

[65] Lionel Robert and Daniel M. Romero. 2015. Crowd Size, Diversity and Performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1379–1382. https://doi.org/10.1145/2702123.2702469

[66] Jeffrey M. Rzeszotarski and Aniket Kittur. 2012. CrowdScape: interactively visualizing user behavior and output. In *The 25th Annual ACM Symposium on User Interface Software and Technology, UIST '12, Cambridge, MA, USA, October 7-10, 2012*. ACM, New York, NY, USA, 55–62. https://doi.org/10.1145/2380116.2380125

[67] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, and Jon Froehlich. 2019. Project Sidewalk: A Web-based Crowdsourcing Tool for Collecting Sidewalk Accessibility Data At Scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. ACM, New York, NY, USA, 62. https://doi.org/10.1145/3290605.3300292

[68] Eric Schwitzgebel. 2011. Belief. In *The Routledge Companion to Epistemology*. Routledge, 40–50.

[69] Yang Shi, Krupal Shah, Wengran Wang, Samiha Marwan, Poorvaja Penmetsa, and Thomas Price. 2021. Toward Semi-Automatic Misconception Discovery Using Code Embeddings. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (Irvine, CA, USA) *(LAK21)*. Association for Computing Machinery, New York, NY, USA, 606–612. https://doi.org/10.1145/3448139.3448205

[70] Simon Singh and Edzard Ernst. 2008. *Trick or treatment: The undeniable facts about alternative medicine.* WW Norton & Company.

[71] Psychology Today. 2022. Why Don't People Believe in Climate Change? https://www.psychologytoday.com/us/blog/psych-unseen/202204/why-dont-people-believe-in-climate-change. (Accessed on 07/10/2022).

[72] usability.gov. [n. d.]. System Usability Scale (SUS). https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html. (Accessed on 07/10/2022).

[73] Miguel A Vadillo, Helena Matute, and Fernando Blanco. 2013. Fighting the illusion of control: How to make use of cue competition and alternative explanations. *Universitas Psychologica* 12, 1 (2013), 261–270.

[74] Miguel A Vadillo, Serban C Musca, Fernando Blanco, and Helena Matute. 2011. Contrasting cue-density effects in causal and prediction judgments. *Psychonomic Bulletin & Review* 18, 1 (2011), 110–115.

[75] Clare R Walsh and Steven A Sloman. 2004. Revising causal beliefs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 26.

[76] Jun Wang and Klaus Mueller. 2015. The visual causality analyst: An interactive interface for causal reasoning. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 230–239.

[77] Mark E. Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan,

Teógenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A. Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, and Michael S. Bernstein. 2017. Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*. ACM, New York, NY, USA, 1902–1913. https://doi.org/10.1145/2998181.2998234

[78] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in Social Media: Definition, Manipulation, and Detection. *SIGKDD Explor. Newsl.* 21, 2 (nov 2019), 80–90. https://doi.org/10.1145/3373464.3373475

[79] Adrienne Yapo and Joseph Weiss. 2018. Ethical implications of bias in machine learning. (2018).

[80] Ion Yarritu, Helena Matute, and David Luque. 2015. The dark side of cognitive illusions: When an illusory belief interferes with the acquisition of evidence-based knowledge. *British Journal of Psychology* 106, 4 (2015), 597–608.

[81] Ion Yarritu, Helena Matute, and Miguel A Vadillo. 2014. Illusion of control: the role of personal involvement. *Experimental psychology* 61, 1 (2014), 38.

[82] Chi-Hsien (Eric) Yen, Haocong Cheng, Yu-Chun (Grace) Yen, Brian P. Bailey, and Yun Huang. 2021. Narratives + Diagrams: An Integrated Approach for Externalizing and Sharing People's Causal Beliefs. *Proc. ACM Hum. Comput. Interact.* 5, CSCW2 (2021), 444:1–444:27. https://doi.org/10.1145/3479588

[83] Chi-Hsien Yen, Haocong Cheng, Yilin Xia, and Yun Huang. 2023. CrowdIDEA: Blending Crowd Intelligence and Data Analytics to Empower Causal Reasoning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 463, 17 pages. https://doi.org/10.1145/3544548.3581021

[84] Ming Yin, Siddharth Suri, and Mary L. Gray. 2018. Running Out of Time: The Impact and Value of Flexibility in On-Demand Crowdwork. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174004

A FORMATIVE STUDY

A.1 Crowd Workers' Expertise Level Regarding Climate Change

We present the self-reported knowledge and agreement level of the crowd workers regarding various climate change-related attributes and statements in Figure 18. Examples of climate change-related attributes are greenhouse gases, deforestation, and the melting of ice. One example of climate change-related statements is: "Climate change is happening right now". We present the attributes and statements in the supplemental material containing survey questions. We aggregate a particular worker's selected levels and bin them into different knowledge and agreement levels.

In terms of knowledge level, only a small portion (<10%) of the crowd workers did not deem themselves knowledgeable about the climate change-related attributes. On the other hand, the majority of the crowd workers seemed to highly agree with climate change-related statements, which indicates that they can be categorized as climate change believers.

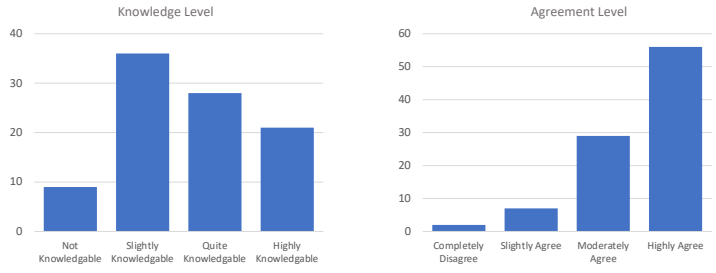

Fig. 18. **Self-reported knowledge and agreement level of the crowd workers regarding climate change.** Y-axes represent counts for each category. The higher agreement level denotes a higher inclination to believe in the existence of climate change.

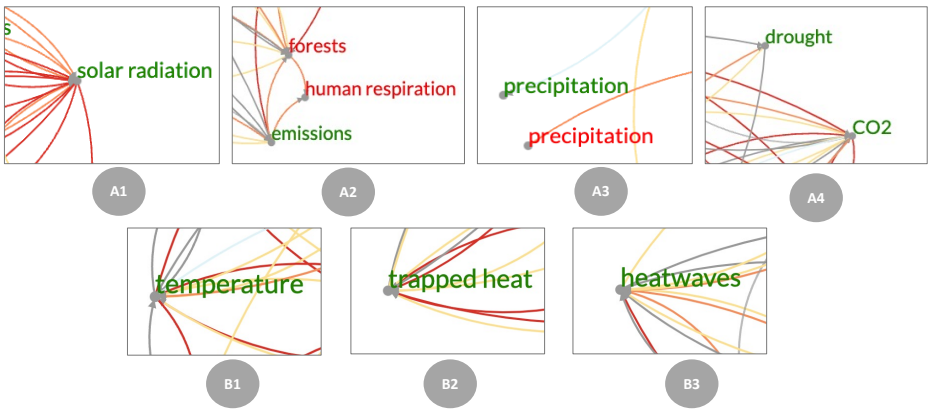

Fig. 19. ***Misinformed* and *Oblivious* cases in discrepancy network.**

A.2 Noteworthy and Interesting Causal Illusion Cases

In Figure 19, we present some noticeable cases within the discrepancy network. A large number of high levels of *misinformed* links (with red and orange links) are related to increasing solar radiation. We previously identified this as a prevalent misconception among climate-change deniers (Figure 19-A1). Less human respiration is an attribute that always appears with links with $cs = 0$ in the ground truth data, but the crowd linked it with decreasing forests and increasing emissions as effects of both (Figure 19-A2). Another interesting finding is that increasing precipitation has more significant or visible links than decreasing precipitation, meaning more people have voted for them. This can be rationalized by the higher ratio of crowd workers from the states of California and Texas (Figure 19-A3). Here it is also interesting that the decreasing precipitation is linked with being misinformed (orange link) while increasing precipitation is linked with being oblivious (lightest blue link).

Some attributes that present some level of visible alignment of the crowd's opinion with the ground truth are more drought and increasing CO2 (a good mixture of red, orange, grey, and blue colored links)(Figure 19-A4). Other various levels of being misinformed are linked with temperature-related attributes such as increasing temperature, more intense heatwaves, and more trapped heat. We think this happened because of a certain ambiguity related to the physical relationships among these three attributes. This validates further refinement of the causal attributes for the next phases of our experiments (Figure 19-B1, B2, B3).

B FINAL STUDY

B.1 Detailed Experimental Protocol

We mention the sequential workflow of the crowd workers below:

(1) **Read the instructions and perform and pass the test.** These tasks were performed in the *Instructions and Overview Module* (Section 5.3.1). Each crowd worker reads an overview of the whole interface as a step-by-step guide, along with the explanation and purpose of each step. They have the opportunity to go back to previous pages or restart anytime. They can refer to the pictures of the interface in different phases of their upcoming workflow that are provided in the module. Next, they encounter a simple instance of building a causal relation between

two attributes. If they can pass this test, they proceed to the next step. If they fail this test, they can always restart this module and retry this test to improve their understanding.

(2) **Complete the demographics survey.** This task was performed in the *Demographics and Climate Change Awareness Survey Module* (Section 8.1.1). Each crowd worker answers 8 questions regarding their demographics, ad 4 questions on climate change awareness. The demographics questions are about their ethnicity, gender, marital status, geographical location (state and county), education, employment status, and age group. In each demographic question, they have an option not to provide their information. We provide the demographics questions in the supplemental materials.

(3) **Create a causal network.** This task was performed in the *Causal Network Creation Module* (Section 8.1.2). Each crowd worker will create five different causal links to build a small causal network. They need to perform two micro-tasks to create each causal link.
- Choose the "cause" (and its respective trend) from the cause drop-down.
- Choose the "effect" (and its respective trend) from the effect drop-down.

The crowd workers are free to choose the attributes and their trends based on their personal perceptions. For example, they may choose the attribute "CO2" with an "increasing" or "decreasing" trend. Performing these two microtasks will automatically create a causal relationship. Say, for the chosen two trended attributes, *increasing emissions* as the "cause" and *increasing CO2* as the "effect", the created causal relation would be *increasing emissions* leads to *increasing CO2*.

After the first link, the workers are asked to select a new node that has not been selected before and an already selected node in order to create a new causal relation, adding to the emerging small network. To ensure that the worker will evolve the graph from an already selected node, the cause (or effect) drop-down menu (whatever the user chooses first) will only show already selected nodes. Let's say, the user has chosen the cause first, where the drop-down had listed only already selected nodes, already part of the evolving graph. In this case, the effect drop-down menu will now show all variables. This will make sure that only a single graph of connected nodes is created.

(4) **Alter causal network.** This task is performed in the *Causal Network Alteration Module* (Section 5.3.4). Each crowd worker can alter their created causal network from the previous step by choosing the link they want to change by left-clicking on it and selecting from the available options for alteration. The options are: *change direction* and *do not modify anything*.

(5) **Evaluate causal network.** This task is performed in the *Causal Network Interpretation and Evaluation Module* (Section 5.3.5). Each crowd worker can read their causal network and view it in a node-link diagram or Directed Acyclic Graph (DAG) format. On a scale of 1 to 5, they also provide their confidence level to the created networks.

(6) **Evaluate the interface**. After creating a causal network, each crowd worker is asked to evaluate the interface based on seven usability and learning statements on a 5-point Likert Scale using the *Interface Evaluation Module* (Section 5.3.6).

(7) **Verification and compensation**. Each crowd worker is provided with a unique code at the end of their participation (on AMT) or redirected to a specific link (on Prolific). We used these mechanisms to validate the results, discard any incomplete data, and compensate the crowd worker.

B.2 Final Aggregated Network Node Exploration Status

We provide the final node exploration status of the aggregated network in Figure 20. The crowd explored all attributes to some extent, except for "decreasing solar radiation" which did not appear in any crowd worker's network.

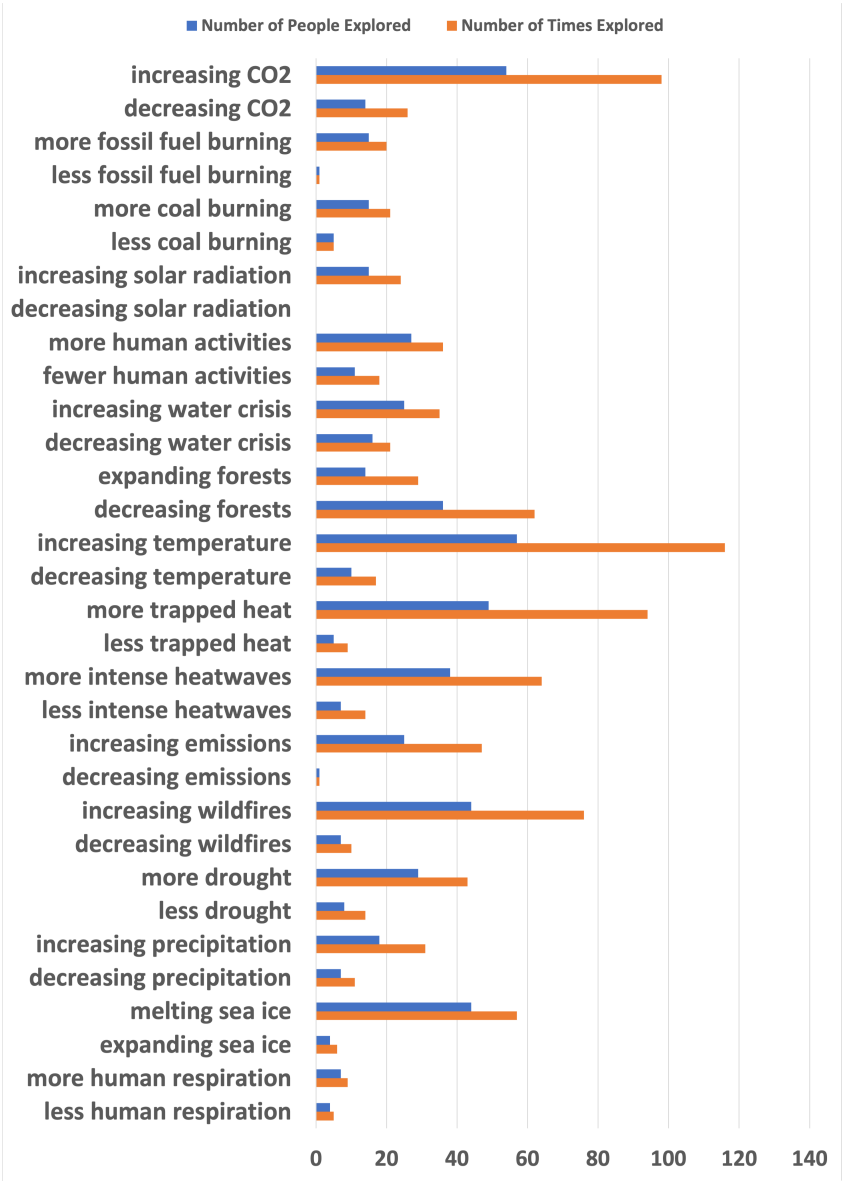

Fig. 20. **The Exploration Status of all nodes denoting the causal attributes in the final aggregated network.** The orange bars denote the number of times the node has been explored either as cause or effect in any causal link created by the crowd. The blue bars denote the number of people who visited and explored that specific causal attribute/node either as cause or effect.

Received January 2023; revised July 2023; accepted November 2023