Extending Counterfactual Reasoning Models to Capture Unconstrained Social Explanations

Stephanie Droop [1] **Neil Bramley** [2]

Abstract

Human explanations are thought to be shaped by counterfactual reasoning but formal accounts of this ability are limited to simple scenarios and fixed response options. In naturalistic or social settings, human explanations are often more creative, involving imputation of hidden causal factors in addition to selection among established causes. Across two experiments, we extend a counterfactual account of explanation to capture how people generate free explanations for an agent's behaviour across a set of scenarios. To do this, we have one group of participants (N=95) make predictions about scenarios that combine short biographies with potential trajectories through a gridworld, using this to crowdsource a causal model of the overall scenario. A separate set of participants (N=49) then reacted to particular outcomes, providing free-text explanations for why the agent moved the way they did. Our final model captures how these free explanations depend on the general situation and specific outcome but also how participants' explanatory strategy is shaped by how surprising or incongruent the behaviour is. Consistent with past work, we find people reason with counterfactuals that stay relatively close to what actually happens, but beyond this, we model how their tendency to impute unobserved factors depends on the degree to which the explanandum is surprising.

1. Introduction

Suppose you see a friend crossing a car park, making a beeline for the end of an occluding wall behind which a food stand is often parked. But your friend stops abruptly at the corner and changes direction. The tradition of Bayesian theory of mind uses the rationality assumption (that people act to achieve their desires given their beliefs) to work backwards to infer agents' beliefs or desires from their behaviour (Baker et al., 2007; 2017; Jara-Ettinger et al., 2020). A salient explanation for this behaviour could be the favourite food stand is absent today. But even if the stand is there, we have no problem coming up with alternative explanations: maybe seeing it reminded your friend of something more urgent she had to do; maybe she felt sick; maybe she saw someone she wanted to avoid. In everyday life, we seem to generate explanations easily and fluently, and readily draw on factors that go beyond the facts given. Natural human behaviour is complex and dynamic, driven by a hierarchy of short- and long-term goals. This presents a challenge for models of social and explanatory reasoning that often depend on a complete pre-existing model, and simplifying assumptions such as that people have stable goals and act in optimal ways to achieve them.

1.1. Explanations and Counterfactuals

A standard account of what it means to explain an event or outcome is to point to preceding event(s) that seem particularly causative for that event's occurrence on this occasion. A person might highlight a lightning strike to explain a fire in a barn over other less unique factors like the presence of hay and oxygen, while a data scientist might explain a model's classification decision on a particular fragment of its input or training data. Either way, explanations involve interrogating one's generative model of the causal relationships between the outcome and the various factors involved in producing it.

A critical component of explanation quality is whether the outcome depends on the highlighted factors not just in the actual world but also across *counterfactuals* — different ways that situation could have played out (Lagnado et al., 2013). Phrased differently, people frequently produce and find satisfying those explanations that pick out variables

[*]Equal contribution [1]Institute for Language, Cognition and Computation, University of Edinburgh, Scotland, United Kingdom [2]Department of Psychology, University of Edinburgh, Scotland, United Kingdom. Correspondence to: Stephanie Droop <stephanie.droop@ed.ac.uk>.

ICML (International Conference on Machine Learning) Workshop on Counterfactuals in Minds and Machines, Honolulu, Hawaii, USA.

which robustly correlate with the outcome across a range of imagined counterfactual scenarios (Quillien, 2020; Gerstenberg et al., 2021). For instance, we more readily blame the lightning than the hay for the barn fire because many things in barns are flammable but in reality rarely catch fire without a spark. Counterfactual accounts which perturb the variables in a situation model to measure the explanatory power of different causes therefore pose a promising account of how people generate explanations. The next section discusses in detail one particular model which we build on in this work.

1.2. Counterfactual Effect Size Model

The *Counterfactual Effect Size Model* (CESM, Quillien & Lucas, 2023) operationalises the notion of simulating variations of what actually happens when selecting causal factors to mention in an explanation. The authors hold that when judging to what extent a cause C explains effect E, people first simulate counterfactual possibilities (as in the structural model pproposed by Lucas & Kemp, 2015), and secondly compute the causal strength of C on E across these counterfactuals.

Important to how this works is the notion of effect size, a measure of correlation of cause with effect across counterfactuals, modelling how reliably an intervention on C would change E on average across a variety of possible background circumstances. Ceteris paribus, the theory is that the more strongly a cause correlates with an effect across counterfactuals, the more likely we are to posit it as an explanation for the effect.

In the CESM, the degree to which simulated counterfactuals depart from what actually happened is controlled by a stability parameter, **s**. When we simulate a counterfactual possibility, for each causal variable in the model, with probability s we leave the variable as it is the actual world. With probability $1 - s$, we instead sample the variable's value from its prior probability distribution, for each counterfactual we can then sample whether the effect occurs or not. Both Lucas & Kemp (2015) and Quillien & Lucas (2023) found that the value of s that maximised correlation with human selections was around 0.7.

The CESM works excellently for predicting explanations for outcomes in simple urn problems (e.g., "If I need two coloured balls to win, to what extent was drawing a blue ball from Urn 1 responsible for my win?", etc., Quillien (2020)). The counterfactual effect size concept has also been shown to give a good account of people's estimates of how causative different states' results were for the overall 2020 US election outcome (Quillien & Barlev, 2022). However, like many probabilistic models of cognitive processes, the CESM assumes all possible explanans are enumerated from the start, leaving the cognizer to simply select which to point at. To make a start towards modelling more naturalistic explanations, such as explanations of the behaviour of other agents ("social explanation"), we investigated the free text explanations people generate when asked to explain the behaviours of agents in settings where there are a range of potentially causal variables to latch onto.

1.3. Our Approach

We investigated social explanation using a novel paradigm where participants react to scenarios involving an agent with four salient personal and situational features, taking one of two trajectories to one of two food stands. We designed the scenarios to vary from over-determined (several variables are salient explanations for their behaviour), through singly determined (one good reason) to surprising or incongruous behaviour (no good reasons for, and several reasons against). We allowed participants to explain the agent's behaviour in each scenario using their own words and then developed a coding scheme to categorise the different explanations people gave. In this way, we explore how people generate explanations in a relatively unconstrained setting.

We are not only concerned with which and how many of the situational factors participants mention, but also with whether and when they posit additional latent causes not mentioned anywhere in the scenario. To model the human-coded responses, we implement CESM (Quillien & Barlev, 2022; Quillien & Lucas, 2023) based on a crowdsourced causal model of the general relationships between the variables manipulated in the scenarios. We build on the CESM by 1) basing our causal model on human intuitions about the relevant relationships 2) allowing for interactions between variables and 3) allowing for "other" responses that refer to latent exogenous factors.

2. Experiments

We ran two connected behavioural experiments (Figure 1). The first (Exp.1a) elicited participants' predictions about how likely a character was to take different paths to different food sources, given various biographical and environmental factors. We used these to construct a generative **situation model** that encapsulates laypeople's intuitions of the casual strengths and interactions between each factor. In the second experiment (Exp.1b) participants were shown a subset of the possible combinations of causal factors and path/food choice outcome. Participants provided a free text explanation for why each agent made that choice. We used the situation model derived from Exp.1a to predict what features of the scene participants would cite in their explanations.

2.1. Gridworld

Both experiments used the same "gridworld", a simple graphic showing an agent walking around in a suburban

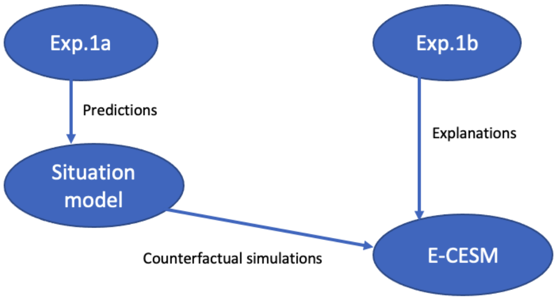

Figure 1: Design and flow of experiment and models.

Table 1: Gridworld Settings

	Factor	Values	Levels
1	Preference	0, 1	Absent, Hotdogs
2	Knowledge	0, 1	Doesn't know area, Knows area
3	Character	0, 1	Lazy, Sporty
4	Start position	0, 1	Hotdog visible, Pizza visible
5	Food choice	0, 1	Pizza, Hotdog
6	Path taken	0, 1	Short, Long

Note: Factors 1:4 describe the situational factors; 5:6 are the agent's choice that participants are asked to explain.

environment before stopping to eat at a food stand. This was accompanied by a short biography text about the agent. We systematically varied three binary biography elements (Preference, Knowledge and Character) and one environmental property (Starting position), yielding 16 scenarios, pairing these with two binary outcome variables (Food and Path), so four potential action outcomes each, or 64 explanation conditions in total (Table 1). Stimuli were the same for both experiments although their presentation was slightly different; see Section 3 and Figure 2 compared to Section 4 and Figure 3. This allowed us to systematically vary different combinations of factors and, first elicit predictions of how likely people found each outcome in each situation (Exp.1a), and second, elicit retrospective explanations for each action in each scenario (Exp.1b).

3. Experiment 1a: Predictions

The aim of Exp.1a was to crowdsource a **situation model**, a representation of people's intuitions of the causal strengths and interactions between factors 1–4 on characters' behaviour (factors 5–6).

3.1. Methods

3.2. Participants

We recruited 90 UK-based participants (42 female, 1 other, age Mean ± sd 40.7 ± 11.5, range 19-66) using the Testable Minds subject pool. They were paid $1.60 and the experiment took Mean ± sd 10.4 ± 4.7 minutes.

3.3. Design

All participants saw all 16 scenarios (specified by Factors 1:4 in Table 1) one by one in a random order. For each trial participants rated the probability of the four possible outcomes. The presentation position on screen of these four was counterbalanced between participants to minimise any left-right bias.

3.4. Stimuli

3.4.1. Biographies

Biography stimuli were eight unique short texts about the agent in the gridworld, varying across three factors (Factors 1:3 in Table 1), each with two levels: *Preference*: {"X's favourite food is hotdog", absent}, *Knowledge*: {"X knows the area well", "X doesn't know the area well"} and *Character*: {"X is sporty", "X is lazy"}. An example is: "Jesse's favourite food is hotdogs. They do not know that area well and are sporty." The agent's name was different for each biography to ensure participants treated each trial and agent independently. Unisex names were used to minimise influence of gender stereotypes.

3.4.2. Gridworld Environment

The gridworld stimuli depicted an agent in a stylised 2D world with houses in the middle and a road around the perimeter. The basic environment was always the same, with a hotdog stand at top right and a pizza stand at bottom left. Three factors were manipulated visually (Factors 4:6 in Table 1): starting location of the agent ({top left, bottom right} aka hotdog visible, pizza visible), and then the two outcome factors depicted with a red arrow: the agent's destination ({hotdog stand, pizza stand}), and the length of the path the agent takes ({long, short}). In Exp.1a, the four possible choices or action outcomes as shown by the red arrow ({short path to hotdog stand, short path to pizza stand, long path to hotdog stand, long path to pizza stand}) were presented all at the same time.

3.5. Procedure

The experiment was implemented in Testable and participants completed it in the browser on their own devices. After calibrating their computer screen, they were presented with the study's information sheet and consent form. See Figure 2

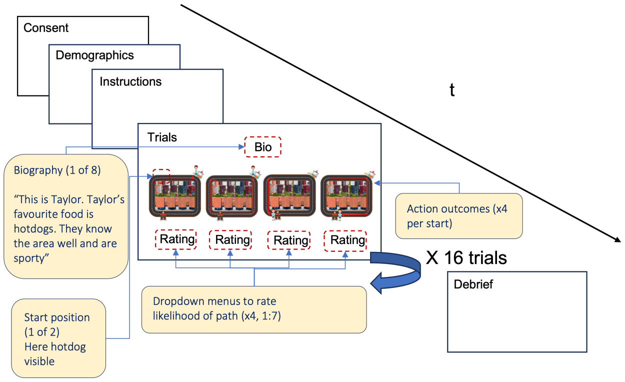
Consent
Demographics
Instructions
t
Trials
Bio
Biography (1 of 8)
"This is Taylor. Taylor's favourite food is hotdogs. They know the area well and are sporty"
Rating
Rating
Rating
Rating
Action outcomes (x4 per start)
X 16 trials
Debrief
Start position (1 of 2) Here hotdog visible
Dropdown menus to rate likelihood of path (x4, 1:7)

Figure 2: Experiment 1a

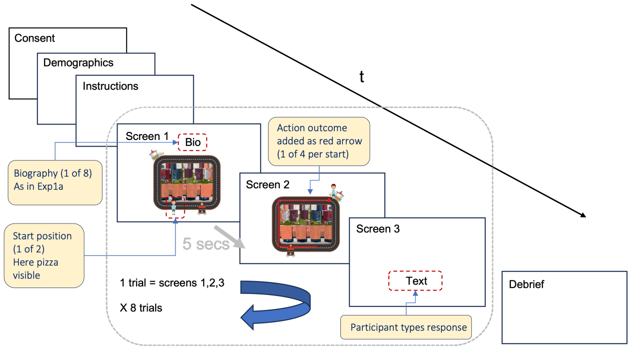
Consent
Demographics
Instructions
t
Screen 1
Bio
Action outcome added as red arrow (1 of 4 per start)
Biography (1 of 8) As in Exp1a
Screen 2
Start position (1 of 2) Here pizza visible
5 secs
Screen 3
1 trial = screens 1,2,3
X 8 trials
Text
Debrief
Participant types response

Figure 3: Experiment 1b

for trial flow schema. Once consent was accepted, participants were given instructions for completing the experiment and shown an example of the stimuli. Importantly, participants were informed that the depicted character could not see through the houses in the middle of the grid-environment. A button was then presented to begin the experiment.

Participants were first shown the agent's biography accompanied by the instruction, "They go for a walk and stop to eat at a food stand. Remember they cannot see through the houses or round a corner. Where do you think they will go? Show how likely each path is by rating each between 1 (not likely at all) and 7 (very likely)". They were then shown each of the four outcomes in counterbalanced order, with a dropdown menu box below each offering the integers 1 to 7. Each participant rated each biography once for each starting position, completing 16 trials in random order.

3.6. Analysis

In Exp.1a, participants had rated how likely from 1 to 7 they thought each of the four possible choices (2 foods × 2 paths), given the character's biography and starting position. These crowdsourced likelihood ratings became the beta slopes of our situation model. To calculate them, we first normalised each participant's ratings to sum to 1 across the four actions for each trial. For example, if they answered 7 for the short path to pizza and 1 for all others, on that trial the short path to pizza was rated 0.7. If they answered that each action was equally likely, then each action was normalised to 0.25, regardless of whether they were all rated 1, 7, or something in between. We then fit two separate generalised logistic mixed-effect regressions, one for each dimension of the action: **food choice** (whether the person went for a pizza or a hotdog) and **path choice** (whether they travelled the shorter or longer way). We included random intercepts for participants. We selected the final model for each dimension using a stepwise procedure implemented by timnewbold/StatisticalModels. By combining the two regressions additively, we obtained the probability of each of the four actions for each situation. This gives rise to the **situation model** which is an intermediate stage in this paper.

3.7. Results: Predictions

Participants in Exp.1a saw each combination of factors in each scenario, and rated how likely was each outcome of food choice and path choice. This means we can use their responses to fit a structural equation model capturing the relationships between the causes and the potential outcomes. Concretely, this **situation model** was a combination of two logistic mixed effects regressions for which we selected the main effects and interaction terms using stepwise model selection. For simplicity, we assumed independent influences of the causes on the choice to take the longer or shorter path and the choice of destination. The resulting model, one outcome of one setting of which is visualised in Figure **??** and another outcome of the same setting in Figure **??** had main effects of Preference, Character and Start Position on food choice, as well as significant interactions, between area Knowledge and Start position, and Character and Start position. Only Knowledge and Character significantly influenced the predicted path length. The resulting model assigns a probability to all four outcomes in each of the 16 scenarios. The odds ratio parameters for each edge can be interpreted straightforwardly as causal influence strengths raising or lowering the probability of the different outcomes. For example, Preference's weight of 5.2 means that, other things being equal, a preference for hotdogs increases the probability of the agent going to the hotdog stand by about a factor of 5.

4. Experiment 1b: Explanations

4.1. Participants

We recruited 49 UK-based adults (40 female, age Mean ± sd 21.1 ± 9.6, range 18-81) using SONA systems online

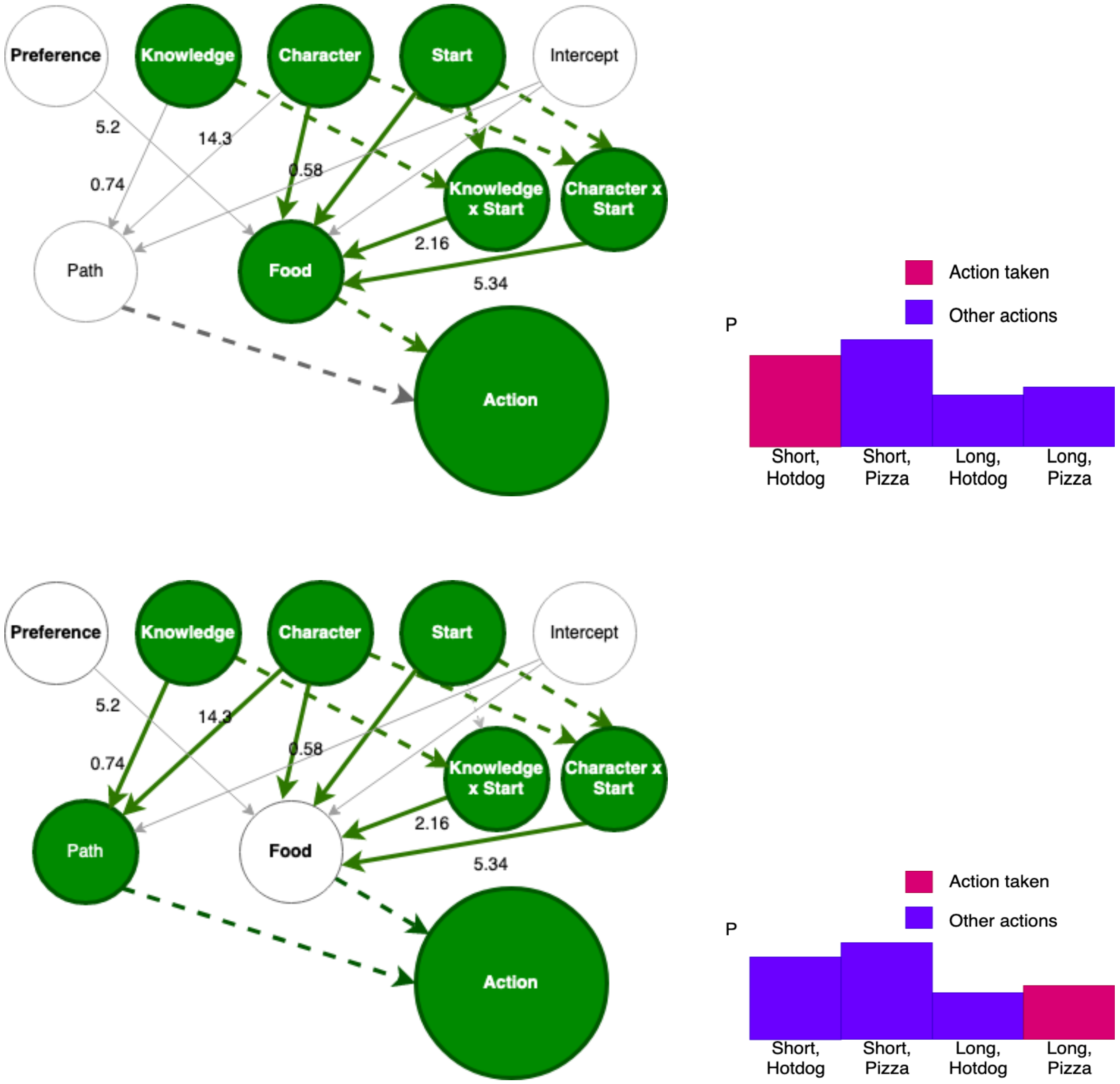

Figure 4: Schema of one setting of the situation model, with two outcomes. Nodes along top left are green for 1 and clear for 0. Here preference is 0, indicating no mention of liking hotdogs. Edge annotations are odds ratios. Solid edges show fitted regression slopes; dotted lines show logical and inferred relationships.

recruitment and a Reddit board for recruiting experimental subjects. Participants were not paid. The task lasted Mean ± sd 10.4 ± 7.9 minutes.

4.2. Design

The 64 stimuli were split into eight groups of eight using a pseudo Latin Square approach. Each participant thus saw one of each of the eight grid configurations and one of each of the eight biographies, but across the sample as a whole all combinations of biography and grid configuration appeared a roughly equal number of times.

4.3. Stimuli

Stimuli were the same gridworlds as Exp.1a.

4.4. Procedure

As per Exp.1a, until presentation of stimuli. Thereafter as per Figure 3. Participants were first shown the agent's biography and starting position in the grid environment. After a few seconds, a red arrow was added to show the agent's choice. At the same time, a text box appeared with the following question written above it: "What do you think is the single best explanation for the person's chosen path?". Once the participant typed their answer, they were presented with another trial with a different stimulus. Each participant saw eight separate trials.

4.5. Analysis

Data were analysed using R version 4.1.

4.5.1. Textual Analysis

Free text responses were stripped of participant and trial data and coded by a research assistant naive to the experiment, by placing a "1" in the relevant column, for whether the participant cited explanations from the biography and situation. The categories were: the agent's Preference (e.g. "They got a hotdog because they like hotdog"), Knowledge of the area, Character (e.g. "They went the long way because they are sporty so they probably wanted a walk before dinner"), their Starting position (a particular food stand was either closer or within sight), or Other which ranged from personal ("He just wanted hotdog today") to situational ("The hotdog stand was closed that day"); see Section 4.7.1 for actual examples. These ratings were then compared to the model predictions after the modelling detailed in the next section.

4.5.2. Extended Counterfactual Effect Size Model ("E-CESM")

We adapted the CESM to apply to our gridworld setup (Figure 5). To obtain model predictions, for each of the 64 gridworlds we simulated outcomes for 1000 counterfactual worlds, for whom the overlap of causal variable states with the actual scenario was governed by a stability parameter as in Quillien & Lucas (2023). For each simulation, the outcome was sampled according to the probability of that action given by the situation model (Figure **??**), creating j rows of sampled counterfactuals for i columns of causes in matrix CF.

Then we calculated the correlation between each causal variable and the actual outcome across these counterfactual worlds. To do this, we looped over i, comparing the subset of the counterfactuals for which the variable in question is 1 ($CF[C_i == 1]$) to those where $CF[C_i == 0]$ and so matching relative proportion of getting the same effect as was actually obtained in the two subsets of counterfactuals.

We optimised stability parameter s through grid search (it was computationally expensive to optimise directly), generating predictions for the model separately for 19 values of s in steps of .05 from .05 to .95. Additionally, in fitting the model to the participant data from Exp.1b, we directly optimised two additional parameters: a parameter τ_1 controlling the probability of providing an explanation that pertains to something not manipulated explicitly, and a softmax temperature parameter on selection τ_2. We additionally hypothesised that the probability of reaching beyond the provided dimensions could be related to how surprising the actual outcome was (i.e. how hard it was to explain in terms of the provided factors). As such we modelled the probability of an explanation being classed as Other as $\tau_1(1 - P(Outcome))$. These parameters were optimised with Nelder-Mead as implemented by R's `optim` function.

Table 2: Comparison of our Extended-CESM model (top) with the others explained in Section 4.5.3.

Model	τ_1	τ_2	s	NLL	BIC
E-CESM	.364	.168	.7	496.4	1010.7
CESM	.228	.127	.8	497.2	1012.3
DD	1.05	1.01	-	545.8	1103.5
Baseline	-	-	-	630.9	1261.8

4.5.3. Alternative Models

For comparison, we also ran the same model predictions through a modified function where propensity to cite Other causes was governed by just a flat τ_1 value rather than being modulated by $1 - P(Outcome)$, assigning the same probability to the other category for all explanations. This represents the classic CESM although that has no provision for outside factors. We also implemented a direct dependency model ("DD") where counterfactual dependence was established just one factor at a time; Finally we calculated a baseline fit which is simply the log likelihood of falling into a category at chance (log(1/5)*392). Results and model comparison are shown in Table 2. Code can be found in our Repository.

4.6. Results: Explanations

The results of Exp.1b consisted of free text verbal explanations for why the character in the gridworld scenario acted the way they did. We used our E-CESM to predict what people would likely mention. Our model had a negative log likelihood of 496.4 and a Bayesian Information Criterion of 1010.7. See Table 2 for how this compares with the baseline and lesioned versions, and Figure 6 for how the final model predictions compare to actual participant data.

The stability parameter s fit best at 0.7, indicating that when simulating counterfactuals, variables kept their original value 70% of the time.

4.7. The Importance of "Other"

In some gridworld settings, participants predominantly answered "Other"; these were cases where the character's choice was surprising given their biography and starting position (as can be seen towards the lower right of Figure 6, which is ordered by increasing unexpectedness of the char-

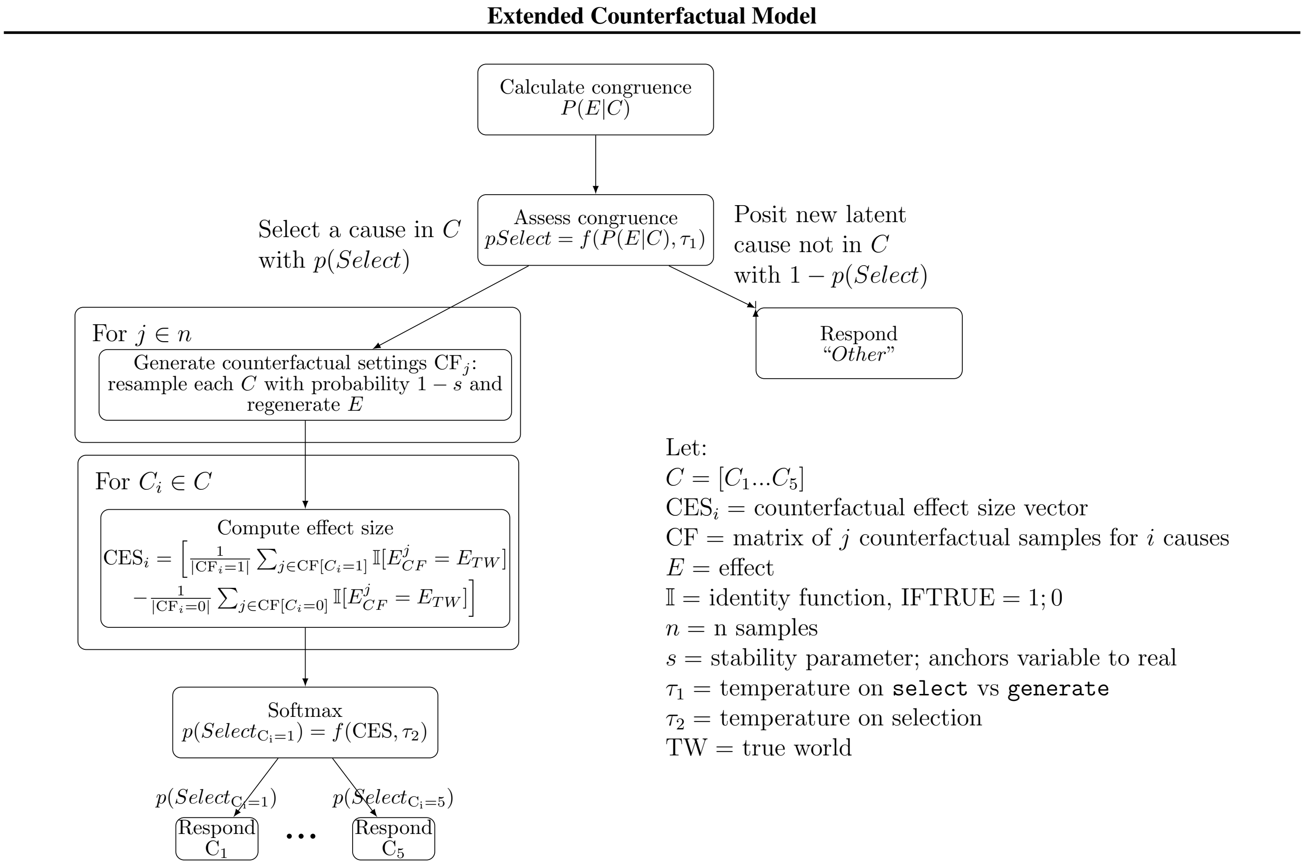
Extended Counterfactual Model
Calculate congruence $P(E|C)$
Assess congruence $pSelect = f(P(E|C), \tau_1)$
Select a cause in C with $p(Select)$
Posit new latent cause not in C with $1 - p(Select)$
Respond "$Other$"
For $j \in n$
Generate counterfactual settings CF_j: resample each C with probability $1 - s$ and regenerate E
For $C_i \in C$
Compute effect size
$\text{CES}_i = \Big[\frac{1}{|\text{CF}_i=1|}\sum_{j\in\text{CF}[C_i=1]}\mathbb{I}[E^j_{CF} = E_{TW}] - \frac{1}{|\text{CF}_i=0|}\sum_{j\in\text{CF}[C_i=0]}\mathbb{I}[E^j_{CF} = E_{TW}]\Big]$
Softmax $p(Select_{\text{C}_\text{i}=1}) = f(\text{CES}, \tau_2)$
$p(Select_{\text{C}_\text{i}=1})$
$p(Select_{\text{C}_\text{i}=5})$
Respond C_1
...
Respond C_5
Let:
$C = [C_1...C_5]$
CES_i = counterfactual effect size vector
CF = matrix of j counterfactual samples for i causes
E = effect
$\mathbb{I}$ = identity function, IFTRUE = 1; 0
n = n samples
s = stability parameter; anchors variable to real
τ_1 = temperature on select vs generate
τ_2 = temperature on selection
TW = true world

Figure 5: Mixture (process) model of E-CESM. The model either selects from available causes (left branch) or generates a new latent cause (right branch) as a function of how surprising (ie. improbable) the action was.

acter's behaviour). For example, the subplot at the bottom right corner represents gridworld 110001, where the character chose the long path to pizza, despite having a preference for hotdog, knowing the area, being lazy, and starting in a position near a visible hotdog stand. Their behaviour could therefore be seen as maximally incongruent with the given facts of the situation and so we would expect both the model and the participants to need to cite Other causes in order to adequately explain the character's choice.

4.7.1. Exploratory Analysis

We performed exploratory qualitative analysis of the "Other" column of the text responses to gain initial insight into any patterns of salient factors. We observed that participants often mentioned temporary changes in the more stable biography factors as well as aspects of the situation. An RA coded each response for mention of temporary *Food state*, *Character state*, *Other state* and *Other situation*.

Out of 392 text responses received, 219 mentioned some kind of Other factor, of which:

- 41 mentioned some kind of temporary desire for a food, e.g. "He had a hotdog recently and wanted a change", "Changed their mind and wanted pizza", "He wanted to catch the aroma of pizza to stimulate his tastebuds before a hotdog".
- 28 mentioned a temporary character state related to those in the biographies but opposed to the current character's biography, e.g. "He was in a lazy mood", "He decided to do exercise for a change".
- 38 mentioned other temporary character states, e.g. "They are tired today", "They changed their mind once they got there".
- 52 mentioned something about the situation outside the person's goals, e.g. "Charlie was procrastinating an assignment", "The hotdog van was closed that day", "They had other things to do in the area", "They got lost".

5. Discussion

In this paper we explored how people explain behaviour in an ecologically richer and more open ended setting than has previously been analysed with experiments and causal models. One way our setup differs from many past scenarios set

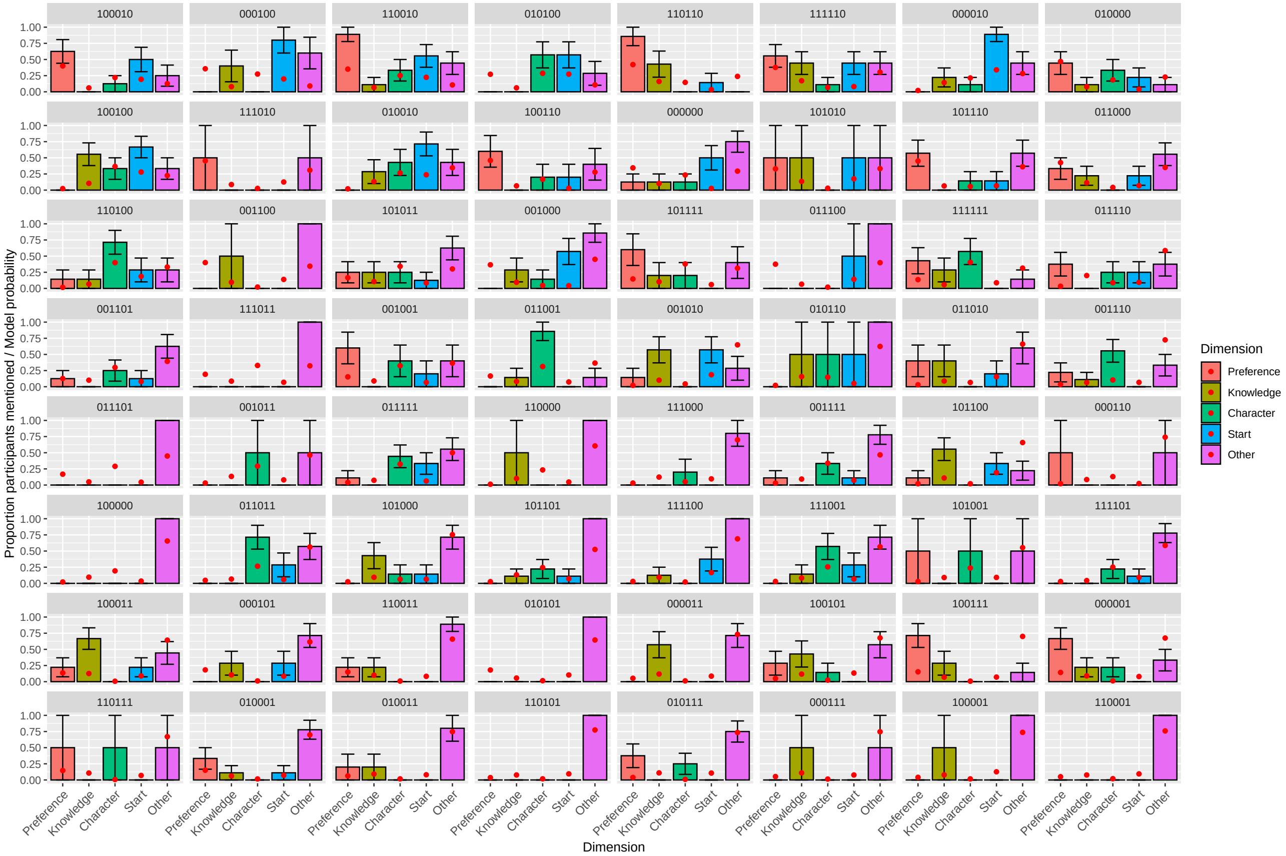

Figure 6: Model predictions (red dots) against participant ratings for each of the 64 gridworld settings, ordered here by the predicted probability of the character's choice. Facet names encode the condition in following sequence: Preference, Knowledge, Character, Start position, Food choice, Path taken following the level conventions in Table 1. For instance the first facet "100010" shows the condition in which the agent likes hotdogs, doesn't know the area, is lazy, can see the hotdog stand, and goes to the hotdog stand by the shorter route.

up is in its coverage: where Lombrozo (2006) and Lucas & Kemp (2015) used scenarios where the behaviour generally made sense under the intended causal situation model, we presented people with a "fully balanced" set of scenarios where all variables were combined with all values of each other.

We first crowdsourced a general model of the situation by asking people to rate how likely the four possible outcomes were for each set of starting values. This revealed that certain behaviours are more or less surprising (why, for example, would a lazy person who has no special preference for hotdogs take the long way round to a hotdog stand they can already see?). Eliciting judgements from people in this way reduces some sources of experimenter-driven assumptions about how people understood the scenario in the task. We then showed new participants each situation-outcome pair, and elicited free explanations. These text responses often made reference to factors from the situation, but also often brought in imaginative reasons from outside the scenario, especially when the behaviour was incongruent. Our extended model could capture that, to some extent, these Other factors tended to dominate the explanations when the behaviour was surprising under the model.

5.1. Comparison with the CESM

We generalise the CESM (Quillien, 2020; Quillien & Lucas, 2023) to a more open-ended setting. Our findings thus offer support for that model and an attempt to bridge the simple and quantified setting of sampling coloured marbles from urns (Quillien & Lucas, 2023) and real world issues like the 2020 US presidential election outcomes (cf. Quillien & Barlev, 2022). Like that study we attempt to bring explanation theories closer to the real world. Unlike it, however, our experimental dataset and modelling is based on human intuitions about the situation rather than a complex statistical model. Their situation model was not proposed to match people's mental models, whereas ours is.

In that light it is noteworthy that the the best fitting stability parameter $s = .70$, is numerically close to the .73 value found in Quillien & Lucas (2023) and .53 in Lucas & Kemp (2015) for their own experimental data and .77 for their reanalysis of Rips (2010). As such there appears to be some converging support for the idea that counterfactuals humans entertain involve resampling causal variables around a third of the time.

5.2. "Other" Causes

Our results also demonstrate that people are rarely unable to generate explanations, even for ostensibly unlikely or surprising behaviours. This shows everyday explanations are considerably richer and more creative than they might appear in tasks that fix the response options to a set of provided causes (e.g., Lombrozo, 2006; Pacer & Lombrozo, 2017). Our results mesh with research suggesting people both tend to overspecify causal relationships when explaining things, and often prefer comprehensive, overdetermined explanations (Zemla et al., 2017), i.e. referring to far more variables than necessary. Although every Bayes net has to simplify its corner of the world and draw artificial lines around the boundaries of a causal system, in reality no system is closed, and people are sensitive to this and able to cast around outside a presented option set.

Our exploratory text analysis suggests social explanations often reference a mixture of individual factors (e.g., personality, preference, etc.) and situational factors (e.g., environmental affordances, distance, convenience, etc.). This brings to mind the long history in social psychology of theories that seek to explain behaviour by making a distinction between *dispositional* factors (those internal to an agent, e.g. ability, knowledge, goals) and *situational* factors (outside the agent's control, e.g., environment, societal pressure) (Heider, 1958/2013) in addition to the later *fundamental attribution error* (Ross, 1977) and *correspondence bias* (Gilbert & Malone, 1995) where people cite situational factors for their own failures, apparently unwilling to concede they may have acted irrationally, but happily cite character or disposition for incongruent behaviour in others. While the CESM does not make any predictions about which would take precedence, and our study was not set up to compare rates at which people are subject to the fundamental attribution bias, exploring self-other differences in what variables are selected in explanations is an avenue for future work.

5.3. Limitations

Limitations to this approach include our simplifying choice to treat the two components of participants behaviour as causally independent, and the relatively small sample size for Exp.1b. Once more data is collected, some of the noise seen between the model predictions and the participant data in Figure 6 should dissipate. We acknowledge the age and gender imbalance between the participant samples of Exp1a and Exp1b, due to 1b being mostly undergraduates and unpaid, but this type of higher-level cognition is not known to have any age or gender differences. Finally and most importantly, although E-CESM is a step towards a model of higher-level cognition, it still only predicts a single "catch-all" Other category rather than truly modelling the flexibility and dynamism of human cognitive processes. However, modelling is in progress toward a generative explanation model which is able to impute hidden variables, similar to the edge replacement technique of Buchanan et al. (2010); Buchanan & Sobel (2014).

5.4. Conclusion

In this paper we presented a computational model of how people explain more or less surprising behaviour. This involved a small extension to an existing counterfactual model of causal selection, enabling it to cover the content of natural language explanations in a naturalistic setting. We combined this with an open-ended free text response format to obtain a richer view of spontaneous explanation, in addition to crowdsourcing a situation model, thereby minimising our need for experimenter-set parameters. Our extension to modulate Other by the probability of the outcome fit provided a modest but encouraging improvement in fit over CESM, a Direct Dependency counterfactual model and a Baseline, making it a promising start toward richer models of human explanation.

References

Baker, C. L., Tenenbaum, J. B., and Saxe, R. R. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10, 2017.

Buchanan, D. and Sobel, D. Edge replacement and minimality as models of causal inference in children. In *Advances in child development and behavior*, volume 46, pp. 183–213. Elsevier, 2014.

Buchanan, D., Tenenbaum, J., and Sobel, D. Edge replacement and nonindependence in causation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 2021.

Gilbert, D. T. and Malone, P. S. The correspondence bias. *Psychological bulletin*, 117(1):21, 1995.

Heider, F. *The psychology of interpersonal relations*. Psychology Press, 1958/2013.

Jara-Ettinger, J., Schulz, L. E., and Tenenbaum, J. B. The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123: 101334, 2020.

Lagnado, D. A., Gerstenberg, T., and Zultan, R. Causal responsibility and counterfactuals. *Cognitive science*, 37 (6):1036–1073, 2013.

Lombrozo, T. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Lucas, C. G. and Kemp, C. An improved probabilistic account of counterfactual reasoning. *Psychological review*, 122(4):700, 2015.

Pacer, M. and Lombrozo, T. Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, 146(12):1761, 2017.

Quillien, T. When do we think that x caused y? *Cognition*, 205:104410, 2020.

Quillien, T. and Barlev, M. Causal judgment in the wild: evidence from the 2020 us presidential election. *Cognitive Science*, 46(2):e13101, 2022.

Quillien, T. and Lucas, C. G. Counterfactuals and the logic of causal selection. *Psychological Review*, 2023.

Rips, L. J. Two causal theories of counterfactual conditionals. *Cognitive science*, 34(2):175–221, 2010.

Ross, L. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*, volume 10, pp. 173–220. Elsevier, 1977.

Zemla, J. C., Sloman, S., Bechlivanidis, C., and Lagnado, D. A. Evaluating everyday explanations. *Psychonomic bulletin & review*, 24:1488–1500, 2017.