Supplementary Information for: ‘Counterfactuals and the logic of causal selection’

Contents

Functional causal models

In this work we use the formalism of Functional Causal Models (FCMs, sometimes also called Structural Equations Models). Here we provide an informal introduction to FCMs; see Pearl, 2000 for a technical treatment.

An FCM represents the world in terms of variables, and the causal connections between them. Consider for example a simple model of whether the pavement gets wet, as a function of whether it rains, and whether the sprinkler is on. We have three variables: R represents whether it is raining, S represents whether the sprinkler is on, and P represents whether the pavement is wet. Each variable can take value 0 or 1 (although note that one can in principle consider FCMs with continuous variables). P is an *endogenous* variable: its value is determined by a functional equation, which specifies that the pavement is wet if it rains or the sprinkler is on:

$$P \coloneqq R \vee S$$

The $\coloneqq$ sign is an assignment operator. Unlike a normal equality sign, it indicates an asymmetry in the equation: the state of the pavement causally depends on the rain and the sprinkler, but not the other way around.

The factors that affect whether it rains, and whether the sprinkler is on, are not explicitly modeled. Instead we sample the values of R and S stochastically when we simulate the system. As such, we say that R and S are *exogenous* variables. Their value is determined by a prior probability distribution. For example, if we live in an area with very dry weather we can set $Pr(R) = .05$, to represent the fact that it rains rarely.

An *intervention* on a variable is the act of setting this variable to the value of our choice, effectively disconnecting it from the causal influence of other variables. The do-operator is a piece of formal notation that allows us to reason about interventions (Pearl, 2000). The term $do(X = 1)$ denotes an intervention that sets the variable X to 1. Using the do-operator, we can coherently talk about probabilities involving the consequences of interventions. For example, $Pr(Y = 1|do(X = 1))$ is the probability of $Y = 1$, given that we made an intervention to set $X = 1$; in some cases this value is different than $Pr(Y = 1|X = 1)$, the probability of Y given that we observe X. For instance, suppose we make an intervention that makes the pavement wet (e.g. by pouring a bucket of water on the ground). In that case, the fact that the pavement is wet does not tell us anything about whether it rained; by contrast, in the absence of an intervention, observing a wet pavement increases our subjective probability that there has been rain; as such, $Pr(R = 1|do(P = 1))$ is different than $Pr(R = 1|P = 1)$.

Necessity-Sufficiency model judgments

Here we give the analytical derivations for the NSM judgments in studies 2-4.

Remember that according to the NSM, the causal strength of an event C for outcome E is:

$$(1 - SP(C))N(C \to E) + SP(C)S(C \to E)$$

where SP(C) is the sampling propensity of C, $N(C \to E)$ is the necessity of C for E, which is 1 if and only if C was necessary for E in the actual world, and $S(C \to E)$ is the sufficiency of C for E, which is $Pr(E = 1|do(C), C = 0, E = 0)$.

Study 2

To compute the sufficiency of event X, we first assume that $X = 0$. Among the worlds where the player loses the game (worlds where his score is less than 2), we compute the proportion of worlds where the player's score is 1 (because these are the worlds where an intervention setting X to 1 would make the player win the game). Therefore, we compute the sufficiency of X as:

$$S(X) = Pr(\text{Score} = 1|\text{Score} < 2)$$
$$= \frac{Pr(\text{Score} = 1)}{Pr(\text{Score} < 2)}$$

These probabilities depend on the sampling propensities of the other two events Y and Z, which we denote $SP(Y)$ and $SP(Z)$. We have:

$$Pr(\text{Score} = 1) = Pr((Y, \neg Z) \vee (\neg Y, Z))$$

$$= SP(Y)SP(\neg Z) + SP(\neg Y)SP(Z)$$

and

$$Pr(\text{Score} < 2) = 1 - Pr(\text{Score} \geq 2)$$

$$= 1 - Pr(Score = 2)$$

$$= 1 - Pr(Y, Z)$$

$$= 1 - SP(Y)SP(Z)$$

Therefore we have

$$S(X) = \frac{SP(Y)SP(\neg Z) + SP(\neg Y)SP(Z)}{1 - SP(Y)SP(Z)}$$

where the sampling probability of a variable V is $\mathrm{SP}(V) = s\delta(V) + (1-s)Pr(V)$, with $\delta(V) = 1$ if V happens in the actual world, and 0 otherwise.

We will refer to the successful draws from the low-, intermediate- and high-probability urns as events L, I and H, respectively.

In study 2a, none of the three events are necessary for the outcome, therefore their causal strength is given by $SP(X)S(X)$, where $SP(X) = s + (1-s)Pr(X)$.

In study 2b, since only L and I happen in the actual world, we have $SP(L) = s + (1-s)Pr(L)$, $SP(I) = s + (1-s)Pr(I)$ and $SP(H) = (1-s)Pr(H)$. Since both L and I were necessary for the outcome, the causal strength of an event $X \in \{L, I\}$ is given by $1 - SP(X) + SP(X)S(X)$.

Study 3

We will refer to the successful draw from the orange urn as R, to the successful draws from the low-probability and high-probability purple urns as P_l and P_h, and to the player winning the game as E.

Sufficiency of R

If $R = 0$ the player always loses the game, so the sufficiency of R is simply equal to the probability that the player draws at least one purple ball:

$$S(R) = SP(P_l \vee P_h) = SP(P_l) + SP(P_h) - SP(P_l)SP(P_h)$$

Sufficiency of P_l ***and*** P_h

Assume $P_l = 0$. Then there are two kinds of situations where $E = 0$:

-when $R = 0$

-when $R = 1$ and $P_h = 0$

An intervention on P_l only makes the player win in the second kind of situation. Therefore we have:

$$S(P_l) = \frac{SP(R)SP(\neg P_h)}{P(\neg E)}$$

$$= \frac{SP(R)SP(\neg P_h)}{SP(R)SP(\neg P_h) + SP(\neg R)}$$

Similarly one can show that the sufficiency of P_h is:

$$S(P_h) = \frac{SP(R)SP(\neg P_l)}{SP(R)SP(\neg P_l) + SP(\neg R)}$$

Causal strength scores

In the 'Three-event' condition, only R is necessary for the outcome. Therefore the causal strength of R is $1 - SP(R) + SP(R)S(R)$, and the causal strength of other events P_x is $SP(P_x)S(P_x)$. The sampling propensity of an event X is $SP(X) = s + (1 - s)Pr(X)$.

In the 'Two-event condition', both R and P_l are necessary for the outcome. Thus, the causal strength of an event $X \in \{R, P_l\}$ is $1 - SP(X) + SP(X)S(X)$, and we have $SP(R) = s + (1 - s)Pr(R)$, $SP(P_l) = s + (1 - s)Pr(P_l)$ and $SP(P_h) = (1 - s)Pr(P_h)$.

Study 4

We will refer to the event 'flipper A sends the ball to the right' as A. A is sufficient whenever $B = 1$, where B is the event 'the right flipper sends the ball toward a blue bucket'. So we have $S(A) = SP(B)$.

In the 'In' condition, A is not necessary for the outcome, so its causal strength is $SP(A)S(A) = SP(A)SP(B)$. In the 'Out' condition, A is necessary for the outcome, so its causal strength is $1 - SP(A) + SP(A)S(A) = 1 - SP(A)(1 - SP(B))$.

General version of the Counterfactual Effect Size Model

The following algorithm generates a causal score $k_{C \to E}$ quantifying how well C qualifies as a cause of E.

a. Simulate a large number of counterfactual worlds by sampling repeatedly from the causal model of the situation. To simulate a world, first randomly sample the value of the exogenous variables[1], then set the values of the endogenous variables according to the structural equations. For each variable V in the causal system, compute the standard deviation σ_V of the variable value across all simulated worlds.

b. For each world generated that way, simulate a counterfactual 'twin' world by making an intervention on C, which sets C to a new, randomly sampled value. Then the values of the endogenous variables in this twin world are set naturally according to the structural equations.

c. For each pair of worlds thus generated, compute the specific causal effect of C on E by taking the ratio of the change in the value of E to the change in the value of C between the two worlds ($\frac{\Delta_E}{\Delta_C}$), and multiplying this ratio by the standardizing factor $\frac{\sigma_C}{\sigma_E}$.

d. The causal score of C on E is the average of all specific causal effects across all pairs of worlds. Formally, we can denote it as $k_{C \to E}$ and write it as:

$$k_{C \to E} = \frac{\sum_{i=1}^{n} (\frac{\Delta E}{\Delta C})_i}{n} \frac{\sigma_C}{\sigma_E}$$

where n is the number of simulated world pairs.

[1] In principle, the CESM is not committed to a particular model of how we sample exogenous variables; in the main text, we argue that people simulate an exogenous variable V with probability $s\delta(V) + (1 - s)Pr(V)$, where s is a stability parameter, $\delta(V)$ is 1 if V happened in the actual world and 0 otherwise, and $Pr(V)$ is the prior probability of V (Lucas & Kemp, 2015).

Re-analysis of existing data

We computed the predictions of the CESM and the NSM, under the assumption that people sample counterfactual possibilities in the way described by the XSM, for data from existing studies on causal judgments (Morris et al., 2019; O'Neill et al., 2021; Quillien & Barlev, 2022; Zultan et al., 2012; Lagnado et al., 2013). These studies cover a wide range of manipulations, stimuli, and dependent variables. Researchers asked participants to make causal judgments about simple games (Morris et al., 2019; O'Neill et al., 2021), real-world events (Quillien & Barlev, 2022), and collaborative tasks (Zultan et al., 2012; Lagnado et al., 2013). Participants were asked how much they agree that a given event caused an outcome (Morris et al., 2019; Quillien & Barlev, 2022), to what degree they think a given event caused another (O'Neill et al., 2021), or how much someone is to blame (Zultan et al., 2012) or is responsible (Lagnado et al., 2013) for an outcome.

Morris et al. (2019) and O'Neill et al. (2021) elicited causal judgments for conjunctive structures (where two causes are jointly necessary for the outcome) and disjunctive structures (where each of two causes would have been individually sufficient for the outcome), while independently manipulating the prior probabilities of the two causes of the outcome in .1 increments.[2]

Quillien and Barlev (2022) asked participants, for each US state that Joe Biden had won during the 2020 U.S. presidential election, to which extent winning that state caused Biden to win the election. They used simulations from two different election forecasts (FiveThirtyEight and The Economist; Silver, 2020; Heidemanns et al., 2020) to generate predictions for the CESM and the NSM.

Zultan et al. (2012) and Lagnado et al. (2013) familiarized participants with a

[2] O'Neill et al. (2021) also asked people to rate their confidence in the causal judgments they had just made. Here we only analyze the causal judgments themselves, not the associated confidence judgments – because modeling those would require additional assumptions about how people generate metacognitive judgments.

collaborative game where each individual in a team can either succeed or fail at a given task. Whether the team wins or fails is a function of the success or failure of each individual in the team. The researchers manipulated the form of this function. For example, in Studies 1 and 2 of Zultan et al. (2012), the team's victory condition is described by the structural equation:

$$V \coloneqq (A \lor B) \land C \land D$$

meaning that the team wins if D wins and C wins and either A or B wins. In study 3, the victory condition was:

$$V \coloneqq ((A \land B) \lor C) \land D$$

(where $\lor$ and $\land$ are symbols for the logical OR and AND).

The researchers also manipulated which players succeeded and failed. They asked participants to what extent a given player was responsible (in Lagnado et al., 2013) for the team's success or failure, or to blame for the team's failure (in Zultan et al., 2012).

To generate model judgments for Morris et al. (2019), O'Neill et al. (2021), Zultan et al. (2012) and Lagnado et al. (2013), we followed the procedure described in Quillien (2020), with the exception that we computed the sampling propensity of a given event X as $SP(X) = s\delta(X) + (1 - s)Pr(X)$, where $\delta(X)$ is 1 if X happens in the actual world, and 0 otherwise. Zultan et al. (2012) and Lagnado et al. (2013) did not give explicit information to their participants about the probability of a given event, so we assumed that each event X had $Pr(X) = .5$ (i.e. each team member in their scenarios had an independent 50% probability of success).

To generate model judgments for Quillien and Barlev (2022), we followed the procedure described in Quillien and Barlev (2022), with the exception that for all simulations in the original forecasting datasets, we randomly forced each state to take its actual-world value with probability s before computing model predictions. For each value of s we independently repeated this procedure 20 times, and we report the average fit between model judgments and average participant judgments across these 20 runs. We tested in this way each value of s between 0 and .98 in .02 increments.

Below we display figures showing the CESM predictions (with the stability parameter set to $s = .73$, its best-fitting values in our new experiments) alongside average human judgments for each study we re-analyzed.

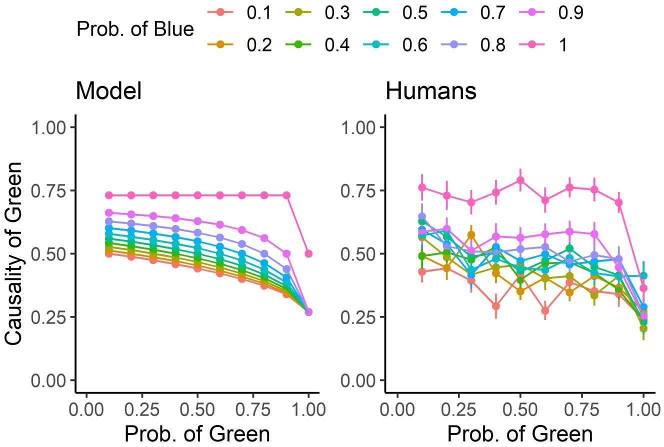

Figure S1

CESM and mean human causal judgments for the Conjunctive structure in Morris et al. (2019).

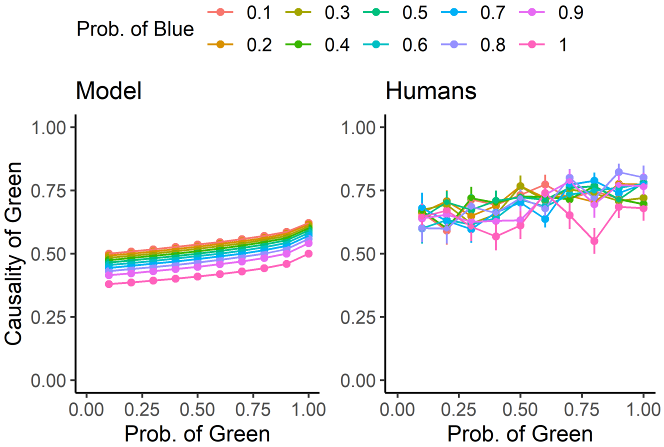

Figure S2

CESM and mean human causal judgments for the Disjunctive structure in Morris et al. (2019).

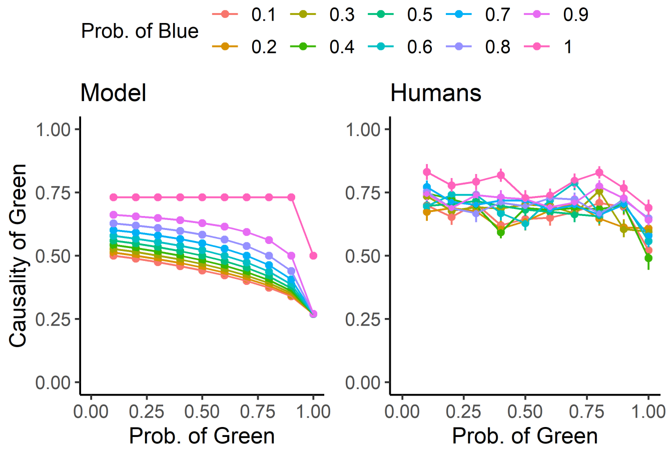

Figure S3

CESM and mean human causal judgments for the Conjunctive structure in O'Neill et al. (2021).

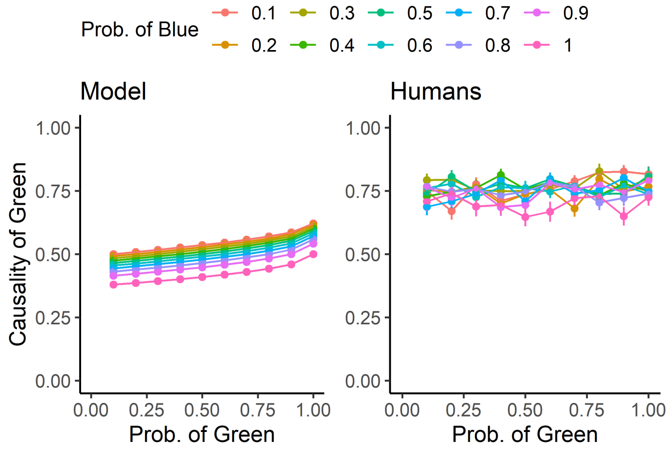

Figure S4

CESM and mean human causal judgments for the Disjunctive structure in O'Neill et al. (2021).

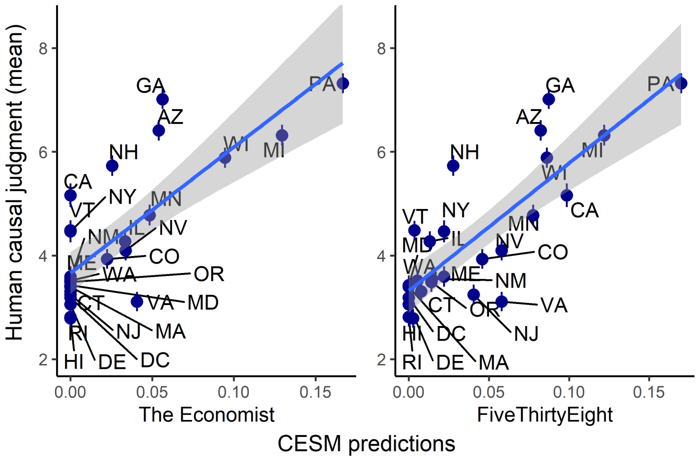

Figure S5

CESM and mean human causal judgments for Quillien and Barlev (2022). Both panels display the same human data, but model predictions are calibrated with simulations from different election forecasts: The Economist (left panel) and FiveThirtyEight (right panel).

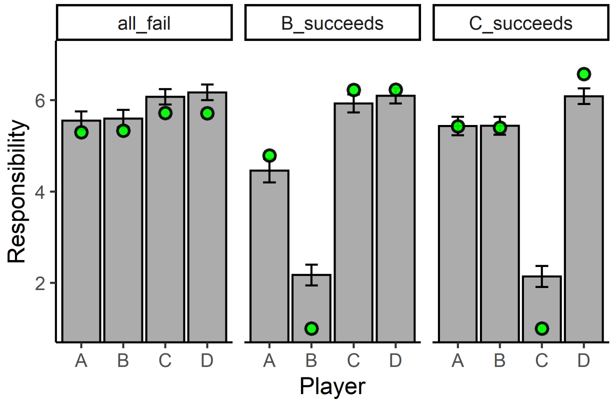

Figure S6

CESM (green) and mean human blame judgments (grey) for Experiment 1 in Zultan et al. (2012). In all conditions, the team fails. all_fail: all players fails. B_succeeds: only player B succeeds. C_succeeds: only player C succeeds. The condition for victory is given by $V := (A \lor B) \land C \land D$.

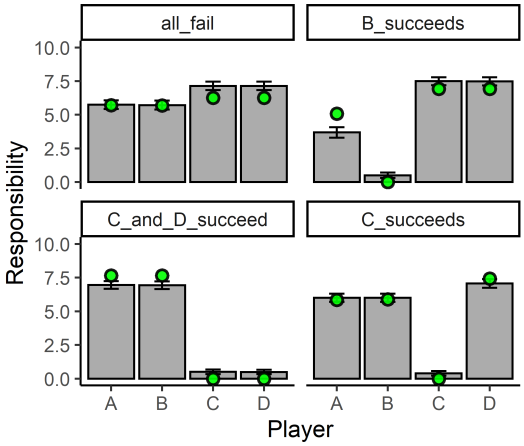

Figure S7

CESM (green) and mean human blame judgments (grey) for Experiment 2 in Zultan et al. (2012). In all conditions, the team fails. all_fail: all players fails. B_succeeds: only player B succeeds. C_and_D_succeed: players C and D succeed. C_succeeds: only player C succeeds. The condition for victory is given by $V := (A \vee B) \wedge C \wedge D$*.*

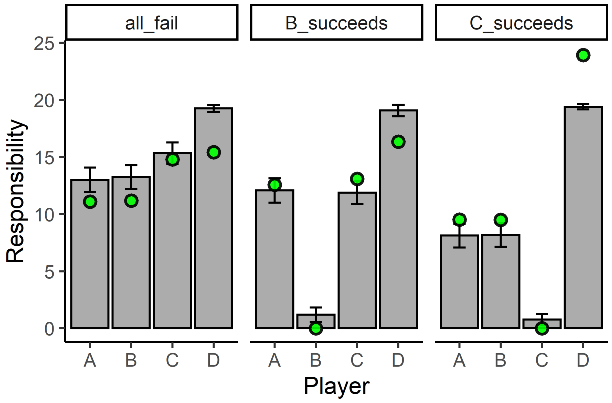

Figure S8

CESM (green) and mean human blame judgments (grey) for Experiment 3 in Zultan et al. (2012). In all conditions, the team fails. all_fail: all players fails. B_succeeds: only player B succeeds. C_succeeds: only player C succeeds. The condition for victory is given by $V := ((A \wedge B) \vee C) \wedge D$ *.*

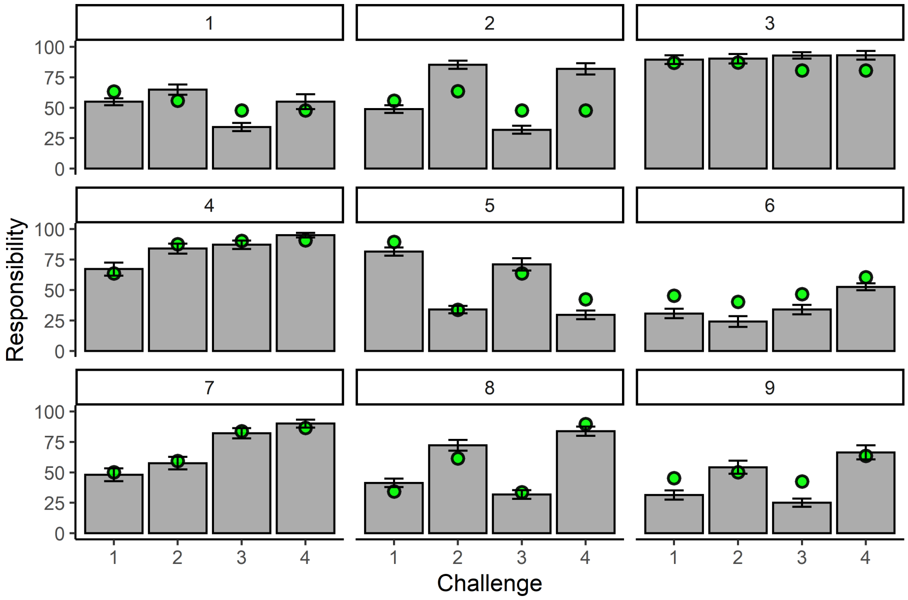

Figure S9

CESM (green) and mean human responsibility judgments (grey) for Lagnado et al. (2013). Each 'challenge' corresponds to one scenario, where participants are asked to judge the responsibility of player A for the team's outcome. A description of each challenge can be found in the .xls datafile for the original paper, available at https://cicl.stanford.edu/publication/lagnado2013causal/

Pivotality and criticality

Here we report a more in-depth analysis of the relationship between our theory and the account developed in Zultan et al. (2012) and Lagnado et al. (2013).

Lagnado and colleagues show that participants' judgments about an individual's responsibility for an outcome seem to track two distinct quantities called the pivotality (Chockler & Halpern, 2004) and the criticality of the individual with respect to the outcome of interest. Here we discuss the meaning of these constructs in the context of the experimental paradigm in Zultan et al. (2012) and Lagnado et al. (2013). As a reminder, the authors familiarized participants with a collaborative game where each individual in a team can either succeed or fail at a given task. Whether the team wins or fails is a function of the success or failure of each individual in the team. Participants are then shown scenarios containing information about which team members failed or succeeded, and whether the team succeeded, and are asked to what extent a given team member is responsible for the team's outcome (failure or success).

Pivotality can be understood as an elaboration of the notion of actual necessity. A player's performance was *actually necessary* for an outcome if an intervention that changes the player's performance would have been enough to change the outcome, holding the performance of all other players constant. Pivotality measures how close a player was from actual necessity. If we denote k the minimal number of interventions on other players one would need to perform in order to make the focal individual actually necessary for the outcome, then an individual's pivotality for the outcome is $1/(1+k)$.

Pivotality is a 'retrospective' notion, in that it takes into account both the general causal structure and what actually happened. In contrast, criticality takes only the general causal structure into account. As such it is a 'prospective' notion, that can be computed even before the actual event. A player A is critical for the team's success to the extent that A's failure would decrease the team's probability of success. For example, if A's success is always required for the team to win, then A has criticality 1. By contrast if B's success can compensate for A's failure, then criticality is 'spread out' between A and B, and each has criticality lower than 1. More formally the criticality of player X is defined as $1 - \frac{Pr(\text{Team Wins}|\text{X fails})}{Pr(\text{Team Wins}|\text{X succeeds})}$.

To better understand the relationship between these two constructs and our model, here we re-plot human and model data (as in Figures S6—S9 above), but with the players' criticality (in blue) and pivotality (in red) added to the figure. Of particular interest are the cases where criticality is held constant while pivotality varies, and vice-versa. These contrasts show that both pivotality and criticality independently contribute to participants' responsibility judgments, and we find that CESM judgments covary with both.

We first illustrate this point by looking at the data from Zultan et al. (2012), see figure

S10.

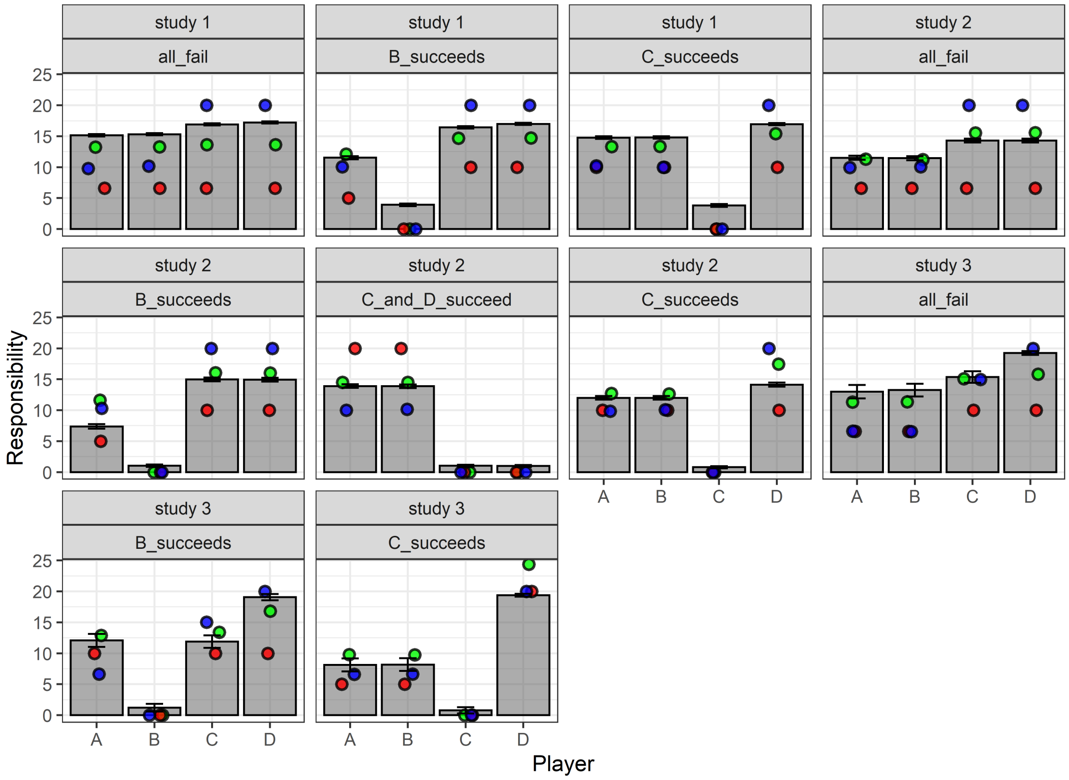

Figure S10

CESM (green) and mean human responsibility judgments (grey) for Zultan et al. (2012), along with the pivotality (in red) and criticality (in blue) of each player. The condition for victory is given by $V := (A \vee B) \wedge C \wedge D$ in Studies 1 and 2, and $V := ((A \wedge B) \vee C) \wedge D$ in Study 3. In all cases, the team loses, and participants rate the responsibility of each player for the team's loss. Error bars represent the standard of the mean.

Consider responsibility judgments for player A in study 3, across the 'B succeeds' and 'C succeeds' cases. Since the causal structure is held constant in study 3 (the condition for victory is always the same), A has the same criticality in both cases. Yet A is more *pivotal* in the 'B succeeds' case, for the following reason. The causal structure in Study 3 is $V := ((A \wedge B) \vee C) \wedge D$; i.e. D's success is always required for a team success, and additionally it must be that either C succeeds or both A and B succeed. In the 'B succeeds' condition, the team would have succeeded if A and D had succeeded, so only one intervention (on D) would have been needed to make the outcome dependent on A. By contrast, in the ‘C succeeds’ condition, A is less pivotal, because one would first need to make C fail and to make B and D succeed in order to make the outcome depend on A. Similarly, both the CESM and human participants assign more responsibility to A in the ‘B succeeds’ relative to the ‘C succeeds’ condition. The CESM assigns higher causal responsibility to pivotal agents, all things being equal, because it preferentially samples counterfactuals that are close to the actual world, and a variable that was pivotal for the outcome tends to be highly correlated with that outcome across nearby possible worlds.

Zultan et al. (2012) also use scenarios where pivotality is held constant while criticality varies. For example in Study 2 in the ‘all fail” condition, every team member has pivotality 1/3, but C and D are more critical than A and B. C and D are also judged more responsible by both participants and the CESM.

We plot the data for Lagnado et al. (2013) in figure S11.

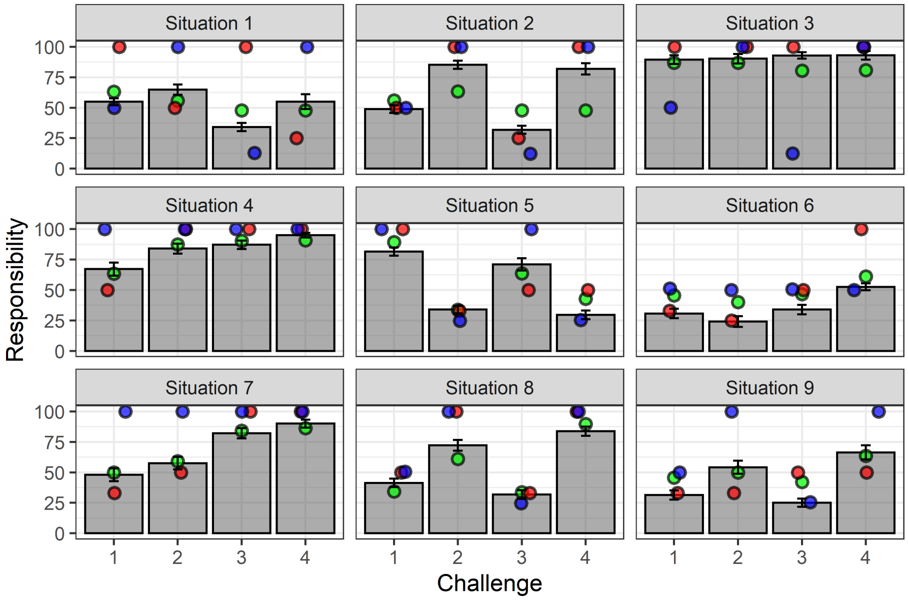

Figure S11

CESM (green) and mean human responsibility judgments (grey) for Lagnado et al. (2013), along with the pivotality (in red) and criticality (in blue) of each player. Each 'challenge' corresponds to one scenario, where participants are asked to judge the responsibility of player A for the team's outcome. A description of each challenge can be found in the .xls datafile for the original paper, available at https://cicl.stanford.edu/publication/lagnado2013causal/. Error bars represent the standard error of the mean.

We also zoom in on Situation 7 and 9 in figures S12 and S13 for illustration. Under each specific scenario we include a diagram representing its causal structure as well as what happened (reproduced from Lagnado et al., 2013). Red crosses and green check marks next to each team member indicate whether that team member has failed or succeeded. Similar symbols next to the trophy indicates whether the team has failed or succeeded. Team success is a function of team members' successes. An arrow from an individual or set of individuals to the trophy indicates that success of at least one individual within the set is necessary (but not necessarily sufficient) for the team to win. For example the causal structure of the scenarios in Situation 7 can be read as: $V \coloneqq A \wedge B \wedge (C \vee D)$.

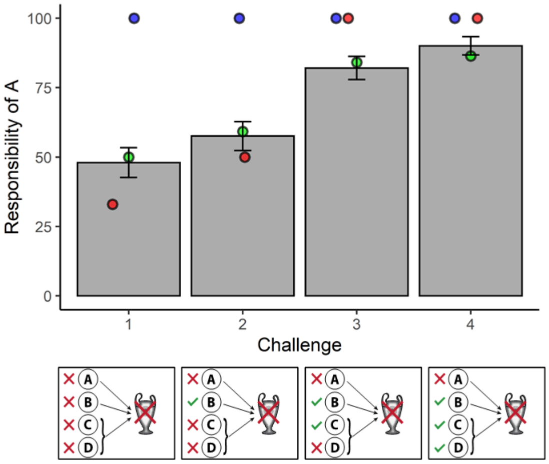

Figure S12

Situation 7 in Lagnado et al. (2013). Mean human responsibility judgments (grey), along with CESM judgments (green), pivotality (red) and criticality (blue). Error bars represent the standard error of the mean.

Across the challenges in Situation 7 (see Figure S12), criticality is held constant (because the causal structure is always the same) while pivotality varies. In all challenges, the team fails, but A's failure is more or less pivotal depending on the challenge. For example, in challenge 2, it would take one intervention (either make C or D succeed) to make A's failure necessary for the team's failure, therefore A has pivotality 1/2. By contrast in challenge 3 A's failure is already necessary for the team's failure, therefore A has pivotality 1. Similarly, both participants and the CESM assign more responsibility to A for the team's failure in challenge 3 compared to challenge 2.

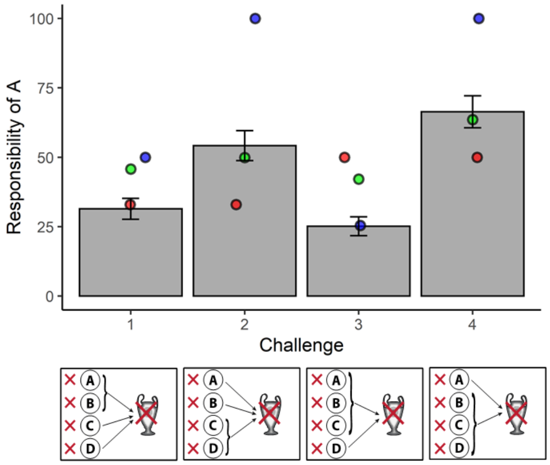

Figure S13

Situation 9 in Lagnado et al. (2013). Mean human responsibility judgments (grey), along with CESM judgments (green), pivotality (red) and criticality (blue). Error bars represent the standard error of the mean.

In situation 9 (see Figure S13), A has the same pivotality in challenges 1 and 2 (in both scenarios it would take two interventions to make A's failure necessary for the team's failure) but A is more critical in challenge 2 than in challenge 1 (in general a failure of player A decreases the team's success probability to a greater extent in challenge 2 than in challenge 1). The same holds for challenges 3 and 4 (A is equally pivotal in each but is more critical in challenge 4). CESM and human judgments tend to covary with criticality (although human judgments do so to a larger extent).

In summary, the CESM tends to reproduce the independent effects of pivotality and criticality on human responsibility judgments, even though it does not explicit represent these constructs.

Alternative counterfactual sampling models

Here we plot the predictions of the CESM, in studies 2-4, under the alternative counterfactual sampling models we describe in the main text. For comparison we also display the human data.

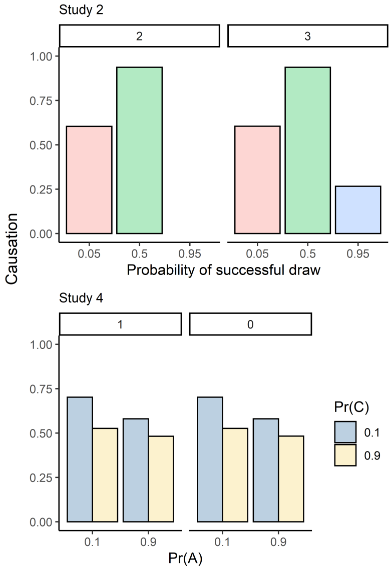

Figure S14

Model predictions for the CESM under the 'whole-world copying' model of counterfactual sampling. Here we used $s = .5$, but the model makes qualitatively identical predictions across values of s.

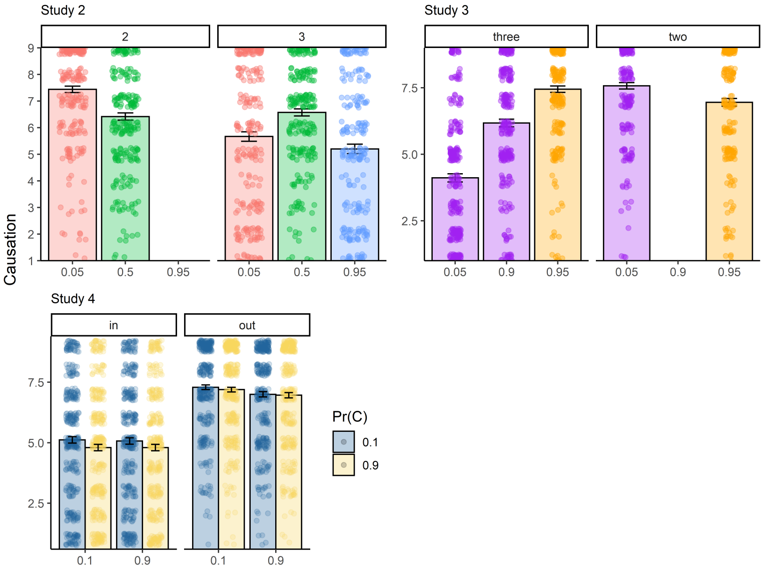

Figure S15

Human causal judgments, studies 2-4.

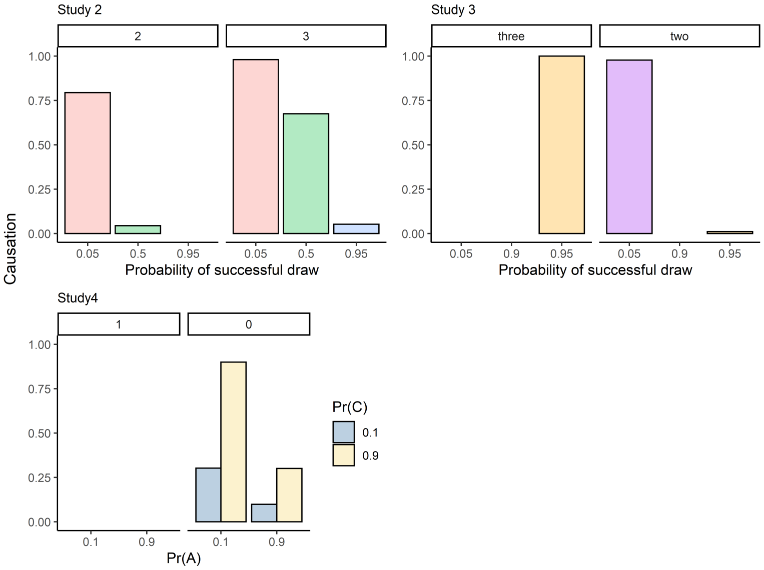

Figure S16

Model predictions for the CESM under the 'minimal deviations' model of counterfactual sampling. Note that this model does not have a stability parameter. In cases where a correlation between two variables is undefined (because one of the variables takes the same value in all minimally deviating worlds), we assume a causal strength of 0 for that variable.

Robustness without probability-raising

In the General Discussion (main text) we discussed the notion of robustness. A causal relationship is not robust if it is heavily moderated by another variable. For instance a drug that cures headaches but only in patients with a certain gene is not a robust cause of headache relief.

In most cases, robust causes raise the probability of their effect more than an non-robust cause would. But one can manipulate robustness while holding probability-raising constant. Consider for example a drug that cures headaches in 25% of patients regardless of their genotype, and contrast it with a drug that cures 50% of patients with allele A and 0% of patients with allele B (we assume that both alleles are equally frequent in the population). The second drug has a less robust effect, but both have the same average effectiveness.

A patient took one of the drugs and was cured as a result. Would people agree more that “the drug caused the patient to be cured" in the condition where the causal relationship is robust (i.e. the drug’s effect is not genotype-dependent)? The predictions of our theory are not straightforward here. On the one hand, our account seems to predict that people will prefer causes that are on average strong but *non*-robust (relative to weaker but robust ones), given that people generate counterfactuals that are centered on the actual world (in most nearby counterfactuals, the patient has the right allele, and so the drug is effective). But pragmatic factors may complicate the situation: people in the ‘non-robust cause’ condition might reason that the patient was cured ‘because he took the drug AND has the right allele’ and find that an explanation that mentions only the drug is pragmatically infelicitous; this would lead them to give lower ratings for the non-robust compared to the robust cause. Relatively little empirical research has looked at such cases, but that research suggests that the direction of the effect depends on the exact wording of the question (see experiment 4 in Grinfeld et al., 2020).

Vasilyeva et al. (2018) also investigate such a case, and find that participants prefer the robust cause, but the token-level causal judgment question they use is not designed to specifically investigate causal selection. Participants observe (for example) a patient who takes the drug and who gets better, but they are not explicitly told whether the patient got better as a result of taking the drug. Therefore the question might be interpreted as a causal inference question, where one has to guess whether the drug had any causal effect on the patient in this particular case (as opposed to the patient coincidentally getting better).

References

Pearl, J. (2000). *Causality*. Cambridge university press.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS one*, *14*(8), e0219704.

O'Neill, K., Henne, P., Pearson, J., & De Brigard, F. (2021). Measuring and modeling confidence in human causal judgment. In *Advances in neural information-processing systems*.

Quillien, T., & Barlev, M. (2022). Causal judgment in the wild: Evidence from the 2020 us presidential election. *Cognitive Science*.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Causality and counterfactuals in group attributions. *Cognition*, *125*(3), 429–440.

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

Silver, N. (2020). How fivethirtyeight's 2020 presidential forecast works—and what's different because of covid-19.

Heidemanns, M., Gelman, A., & Morris, G. E. (2020). An updated dynamic bayesian forecasting model for the us presidential election.

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

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*, 93–115.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*, 1069.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.