Supplementary Information for 'When do we think that X caused Y?'

1 Structural equation models

In order to explain how people assign causes to events, we need to have a model of how the mind represents the causal structure of a situation. To that end, we use the formalism of structural equation models (Pearl, 2000; Halpern, 2016). In a structural equation model, the world is described in terms of variables. For example, the variable F can be used to represent whether the forest is on fire, where F takes the value 0 if the forest is not on fire, and 1 if it is (alternatively, one may treat F as a continuous variable representing the intensity of the fire). A variable is either endogenous or exogenous, depending on whether it is influenced by other variables in the model. The value of an endogenous variable is set deterministically as a function of other variables, as defined by a structural equation. Values of the exogenous variables are set randomly according to probability distributions specific to each such variable.

For example, we can model the causal system described in the forest fire scenario with one endogenous variable F, representing the state of the forest, and three exogenous variables: a variable L representing whether the match is lit, a variable D representing whether the ground is dry, and a variable

Ox representing the amount of oxygen in the air. We assign a probability distribution to the value of each exogenous variable; here the average value of Ox would be high, with a relatively low standard deviation; there would be a low probability that L = 1, and a medium probability that D = 1. The state of the endogenous variable F is determined by the structural equation:

$$F := f(Ox, L, D)$$

where f is some function. In the simple case where we model all three exogenous variables as binary, the equation would be:

$$F := min(Ox, L, D)$$

Such that F takes the value 1 if and only if there is oxygen in the air AND the match is lit AND the ground is dry.

In the structural equation formalism, an *intervention* is defined as the act of setting a variable V to a new value v of one's choice. We do so by replacing the structural equation (or, in the case of an exogenous variable, the probability distribution) that would normally determine the variable's value by a new equation $V = v$. An intervention can be seen as analogous to performing an experiment where we manipulate the value of the variable. The concept of intervention is central to defining causality (Pearl, 2000; Halpern, 2016). Intuitively, if C is to count as a cause of E, there must exist a context in which an intervention on C would result in a change in the value of E (Halpern, 2016).

2 Proof that the causal metric reduces to a correlation coefficient under the no-confounding assumption

Let X and Y be two binary variables (with possible values 0 and 1), which obey the no-confounding assumption; that is:

$$Pr(y|do(x)) = Pr(y|x)$$

for all possible values x and y of X and Y, respectively (Pearl, 2000). $Pr(y|do(x))$ denotes the probability of $Y = y$ given that we have performed an intervention setting X to x (Pearl, 2000).

We show that the causal metric $k_{X \to Y}$ is equivalent to the correlation between X and Y across randomly drawn, independent counterfactual worlds.

We have:

$$k_{X \to Y} = b^k_{Y,X} \frac{\sigma_X}{\sigma_Y}$$

where $b^k_{Y,X} = \frac{\sum_{i=1}^{n^*} (\frac{\Delta Y}{\Delta X})_i}{n^*}$, and n^* is the number of world pairs where $\Delta X \neq 0$.

With binary variables, we can classify the world pairs generated by the sampling-resampling algorithm according to the three following categories:

-world pairs where $\Delta X = 0$,

-world pairs where $\Delta X = 1$,

-world pairs where $\Delta X = -1$

The first kind of world pair is ignored for the computation of $b^k_{Y,X}$, since $\frac{\Delta Y}{\Delta X}$ is undefined when $\Delta X = 0$. We denote the number of world pairs of the second kind as l, and the number of world pairs of the third kind as m.

World pairs of the second kind are made of worlds where $X = 0$ in the original world and $X = 1$ in the twin world. Therefore the proportion of twin worlds with $Y = 1$ is equal to $P(Y = 1|do(X = 1))$, while the proportion of original worlds with $Y = 1$ is equal to $P(Y = 1|X = 0)$. It is easy to see that $\frac{\sum_{i=1}^{l}(\Delta Y)_i}{l}$ must be equal to the difference between the proportion of twin worlds with $Y = 1$ and the proportion of original worlds with $Y = 1$. Therefore we have

$$\frac{\sum_{i=1}^{l}(\Delta Y)_i}{l} = Pr(Y = 1|do(X = 1)) - Pr(Y = 1|X = 0)$$

i.e.

$$\frac{\sum_{i=1}^{l}(\frac{\Delta Y}{\Delta X})_i}{l} = Pr(Y = 1|do(X = 1)) - Pr(Y = 1|X = 0)$$

World pairs of the third kind are made of worlds where $X = 1$ in the original world and $X = 0$ in the twin world. Therefore the proportion of twin worlds with $Y = 1$ is equal to $Pr(Y = 1|do(X = 0))$, while the proportion of original worlds with $Y = 1$ is equal to $Pr(Y = 1|X = 1)$. $\frac{\sum_{i=1}^{m}(\Delta Y)_i}{m}$ is equal to the difference between the proportion of twin worlds with $Y = 1$ and the proportion of original worlds with $Y = 1$. Therefore we have

$$\frac{\sum_{i=1}^{m}(\Delta Y)_i}{m} = Pr(Y = 1|do(X = 0)) - Pr(Y = 1|X = 1)$$

i.e.

$$\frac{\sum_{i=1}^{m}(\frac{\Delta Y}{\Delta X})_i}{l} = -(Pr(Y = 1|do(X = 0)) - Pr(Y = 1|X = 1))$$
$$= Pr(Y = 1|X = 1) - Pr(Y = 1|do(X = 0))$$

By the no-confounding assumption, we have

$$Pr(Y = 1|do(X = 1)) - Pr(Y = 1|X = 0) = Pr(Y = 1|X = 1) - Pr(Y = 1|do(X = 0))$$

$$= Pr(Y = 1|X = 1) - Pr(Y = 1|X = 0)$$

i.e. the average $\frac{\Delta Y}{\Delta X}$ is the same in world pairs of the second and third kinds.

$b^k_{Y,X}$ is the average $\frac{\Delta Y}{\Delta X}$ in the world pairs of the second and third kind:

$$b^k_{Y,X} = Pr(Y = 1|X = 1) - Pr(Y = 1|X = 0)$$

This is simply the regression coefficient $b_{Y,X}$ of Y on X; that is, we have $b^k_{Y,X} = b_{Y,X}$. The correlation coefficient $r_{Y,X}$ between two variables X and Y is related to the regression coefficient $b_{Y,X}$ of Y on X via the following formula:

$$r_{Y,X} = b_{Y,X} \frac{\sigma_X}{\sigma_Y}$$

Therefore we have:

$$r_{Y,X} = b^k_{Y,X} \frac{\sigma_X}{\sigma_Y} = k_{X \to Y}$$

3 Analytical expressions for actual causal strength in simple conjunctive and disjunctive causal structures

Here we derive the analytical expressions for our actual causal strength score, in the simple causal structures used in Morris et al. (Morris, Phillips, Gerstenberg, & Cushman, 2019). The variable G takes the value 1 if Joe draws a green ball from the left box, 0 if he draws a non-green ball from the left box. Similarly, B is the variable representing whether Joe draws a blue ball from the right box, and D represents whether Joe wins a dollar. In both causal structures, the relation between "Joe draws a green ball from the left box" and "Joe wins a dollar" obeys the no-confounding assumption, since there is no variable with a causal influence on them both. Therefore, the causal score $k_{G \to D}$ simply corresponds to the correlation between "Joe draws a green ball from the left box" and "Joe wins a dollar" across sampled counterfactual worlds. Here we compute these correlations in the limit of an infinity of samples.

The correlation coefficient $r_{Y,X}$ between two variables X and Y is related to the regression coefficient $b_{Y,X}$ of Y on X via the following formula:

$$r_{Y,X} = b_{Y,X} \frac{\sigma_X}{\sigma_Y}$$

Therefore we have:

$$r_{D,G} = b_{D,G} \frac{\sigma_G}{\sigma_D}$$

The variance of a random binary variable X is $Pr(X = 1)(1 - Pr(X = 1))$. For conciseness, we will denote probabilities using the shorthands $g = Pr(G = 1)$; $b = Pr(B = 1)$; and $d = Pr(D = 1)$. Therefore we have:

$$\sigma_G = \sqrt{g(1-g)}$$

and

$$\sigma_D = \sqrt{d(1-d)}$$

This yields:

$$r_{D,G} = b_{D,G} \sqrt{\frac{g(1-g)}{d(1-d)}}$$

We will use this along with the fact that, in the case of binary variables, we have:

$$b_{D,G} = Pr(D = 1|G = 1) - Pr(D = 1|G = 0)$$

(intuitively, the regression coefficient quantifies how much the probability that D = 1 changes for each unit change in the value of G).

3.1 Conjunctive causal structure

In the conjunctive causal structure, drawing a green ball from the left box instead of not drawing a green ball (or vice-versa) makes a difference to the outcome only if one has drawn a blue ball from the right box. Therefore we have:

$$b_{D,G} = Pr(D=1|G=1) - Pr(D=1|G=0) = Pr(B=1) = b$$

We also have

$$d = Pr(G=1 \wedge B=1)$$
$$= gb$$

Combining these, we have:

$$r_{D,G} = b\sqrt{\frac{g(1-g)}{gb(1-gb)}}$$

i.e.

$$r_{D,G} = \sqrt{\frac{gb^2(1-g)}{gb(1-gb)}}$$
$$r_{D,G} = \sqrt{\frac{(1-g)b}{(1-gb)}}$$

.

3.2 Disjunctive causal structure

In the disjunctive causal structure, drawing a green ball from the left box instead of not drawing a green ball (or vice-versa) makes a difference to the outcome only if one has *not* drawn a blue ball from the right box. Therefore we have:

$$b_{D,G} = Pr(D=1|G=1) - Pr(D=1|G=0) = 1 - Pr(B=1) = 1 - b$$

We also have

$$d = Pr(G=1 \vee B=1)$$

$$= g + b - gb$$

Combining these, we have:

$$r_{D,G} = (1-b)\sqrt{\frac{g(1-g)}{d(1-d)}}$$

$$= (1-b)\sqrt{\frac{g(1-g)}{(g+b-gb)(1-g-b+gb)}}$$

$$= \sqrt{\frac{(1-b)g(1-b)(1-g)}{(g+b-gb)(1-g-b+gb)}}$$

$$= \sqrt{\frac{(1-b)g(1-g-b+bg)}{(g+b-gb)(1-g-b+gb)}}$$

i.e.

$$r_{D,G} = \sqrt{\frac{(1-b)g}{g+b-gb}}$$

4 Causal strength measure from Morris et al. (2018)

In the case where C and E are binary variables which obey the no-confounding assumption, Morris et al. (2018) derived the following measure for the causal strength of C for E:

$$TC_{C\to E} = \begin{cases} \frac{P(\neg C)P(E)}{P(\neg E)P(C)} & \text{if C was necessary for E} \\ 0 & \text{otherwise} \end{cases}$$

Here we show that, if we make the additional assumption that the causal structure is one where C is necessary for E[1] (i.e. a causal structure where $P(E|\neg C) = 0$), then $TC_{C\to E} = k^2_{C\to E}$.

[1]These are the kinds of causal structures where the model by Morris et al. (2018) has

Earlier we showed that the correlation between two binary variables is:

$$r_{C,E} = [P(E|C) - P(E|\neg C)]\sqrt{\frac{P(C)P(\neg C)}{P(E)P(\neg E)}}$$

In order to relate $r_{C,E}$ to $TC_{C\to E}$, a crucial step is to find a way to express conditional probabilities in terms of elementary probabilities. This can be achieved thanks to the assumption that $P(E|\neg C) = 0$. By the law of total probability, we have:

$$P(E) = P(E|C)P(C) + P(E|\neg C)P(\neg C)$$

since $P(E|\neg C) = 0$, this implies

$$P(E|C) = \frac{P(E)}{P(C)}$$

Therefore the correlation between C and E can be rewritten as:

$$r_{C,E} = \frac{P(E)}{P(C)}\sqrt{\frac{P(C)P(\neg C)}{P(E)P(\neg E)}}$$

$$r_{C,E} = \sqrt{\frac{P(C)P(\neg C)P(E)^2}{P(E)P(\neg E)P(C)^2}}$$

$$r_{C,E} = \sqrt{\frac{P(\neg C)P(E)}{P(\neg E)P(C)}}$$

---

had the most empirical success. As a technical point, note that the 'C was necessary for E' in the definition of $TC_{C\to E}$ is not equivalent to the assumption that the causal structure is one where C is necessary for E. 'C was necessary for E' means that C was necessary in the present situation; this is slightly different from assuming that C is in general necessary for E in the causal structure. For instance, in a disjunctive causal structure, C is not necessary for E in general (i.e. $P(E|\neg C) \neq 0$), although C is necessary for E in the specific situation where C = 1, E = 1, but all other variables are set to 0.

Given the assumption that the causal structure is such that C is necessary for E, then in a situation where $C = 1$ and $E = 1$, C was necessary for E. Therefore we have $TC_{C\to E} = \frac{P(\neg C)P(E)}{P(\neg E)P(C)}$, which leads to:

$$r_{C,E} = \sqrt{TC_{C\to E}}$$

The no-confounding assumption allows us to substitute $k_{C\to E}$ for $r_{C,E}$, yielding:

$$TC_{C\to E} = k^2_{C\to E}$$

5 Non-linear patterns in Morris et al.(2019)

5.1 Conjunctive structure

Both the data and our model show the same nonlinear effects: abnormal inflation is at its strongest when $Pr(green)$ goes from .9 to 1, and supersession is at its strongest when $Pr(blue)$ goes from .9 to 1. This can be easily seen from Figure 1 and 2 in the main text.

5.2 Disjunctive structure

Morris et al. report the three following non-linear effects in the human data for the disjunctive causal structure (see figs. S1 to S3):

a) When the alternate variable is certain ($Pr(blue) = 1$), abnormal deflation ceases to occur

b) When the focal variable is certain ($Pr(green) = 1$), reverse supersession ceases to occur

c) Reverse supersession is at its most powerful as the alternate variable approaches certainty (as $Pr(blue)$ approaches 1).

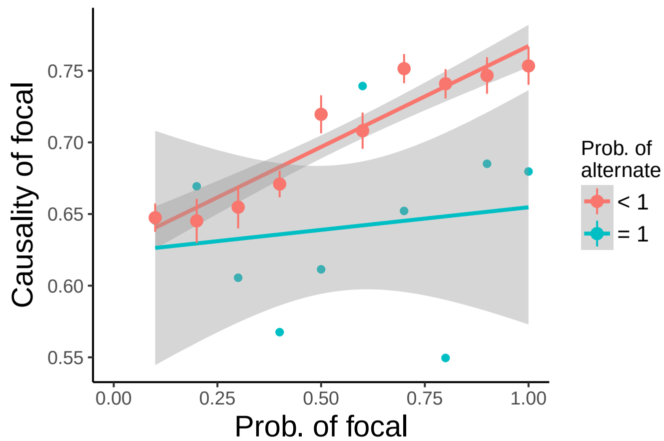

Figure S1: **Effect a.** Abnormal deflation ceases to occur when the alternate is certain.

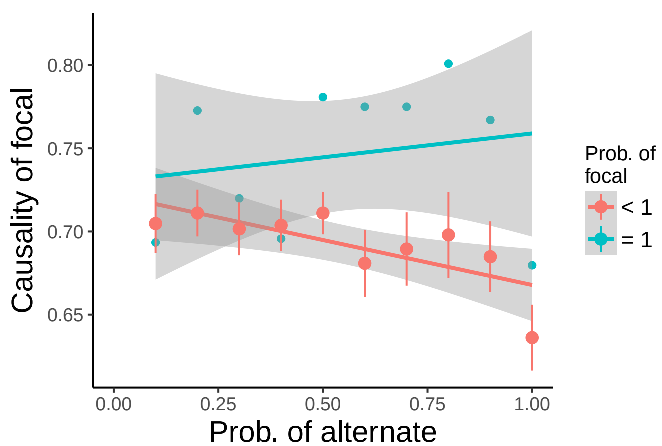

Figure S2: **Effect b.** Reverse supersession ceases to occur when the focal is certain.

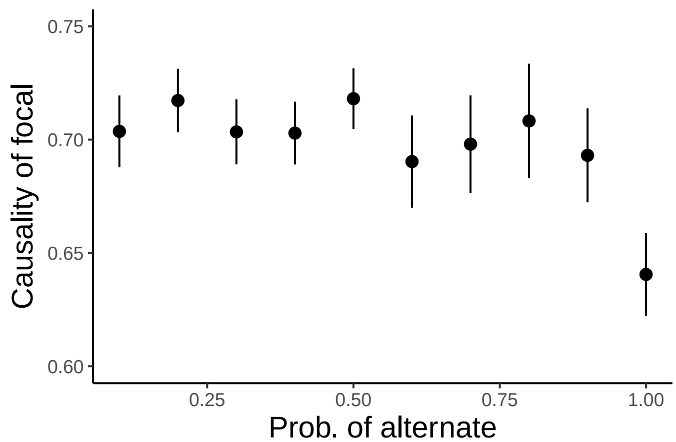

Figure S3: **Effect c.** Reverse supersession occurs mostly as the alternate variable approaches certainty (as $Pr(blue)$ approaches 1).

In what follows, we give an informal discussion of how these effects may be explained by our model, and why they suggest that people use a slightly different normalization procedure than the one we test in the main text.

Effects a) and c) follow naturally from the correlation model:

Explanation of (a): When $Pr(blue) = 1$, then Joe always wins a dollar no matter whether he draws a green ball or not. Therefore, the frequency of green balls in the box does not matter to his likelihood of winning: the correlation between 'Joe draws a green ball from the left box' and 'Joe wins a dollar' across counterfactual worlds is always 0, no matter the value of $Pr(green)$ (except when $Pr(green) = 1$). As a result, abnormal deflation disappears when $Pr(blue) = 1$.

Explanation of (c): When $Pr(blue)$ approaches 1, the causal effect of green drops to 0 (for the reason just outlined above), no matter the value of $Pr(green)$. Therefore, reverse supersession is at its most powerful when $Pr(blue)$ approaches 1.

Effect (b) makes sense if we assume that people normalize the causal effect of green by comparing it to the causal effect of blue. When $Pr(green) = 1$, the correlation between "Joe draws a blue ball from the right box" and "Joe wins a dollar" is 0, so the causal effect of blue is 0, no matter the value of $Pr(blue)$ (except when $Pr(blue) = 1$). When the causal effect of blue is 0, then the normalization procedure always attribute 100% of the overall causal effect to green. As a result, reverse supersession disappears when $Pr(green) = 1$.

In other words, the three effects can be explained by the fact that when one variable is certain, the correlation between the other variable and the outcome is 0, so the unnormalized causal effect of that variable is 0. Indeed, the unnormalized version of our model exhibits effects (a) and (c) (see figs. S4 and S5). However, note that these explanations only work if we assume that the normalization procedure treats a causal effect of 0 as a 'true zero', and does not convert it into another number before comparing it to the causal effect of other variables.

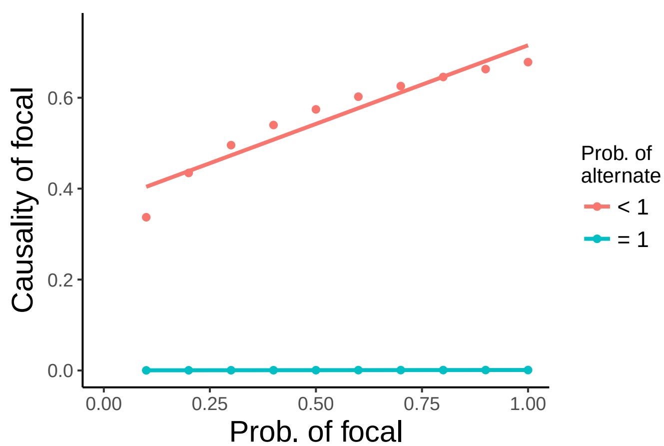

Figure S4: **Effect a in the predictions of the baseline model.** Abnormal deflation ceases to occur when the alternate is certain.

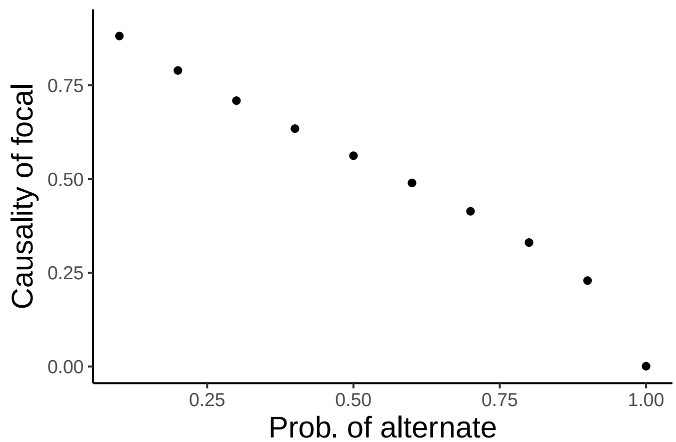

Figure S5: **Effect c in the predictions of the baseline model.** Reverse supersession occurs mostly as the alternate variable approaches certainty.

Yet the normalization we use (following (Morris et al., 2018)) relies on a softmax function, which exponentiates the raw causal effect of each variable before comparing them to each other. By doing so it transforms 0s into 1s (since $e^0 = 1$). As a result, when we examine the normalized version of our model, the three non-linear effects we find in the human data are absent, or weak (see figs. S6 to S8).

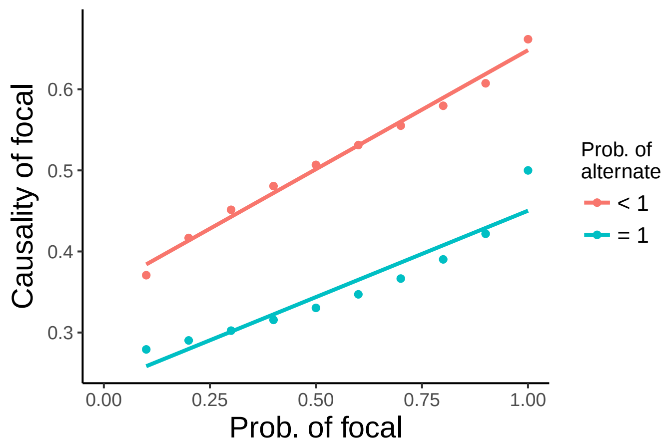

Figure S6: **(absence of) Effect a in the normalized model.**

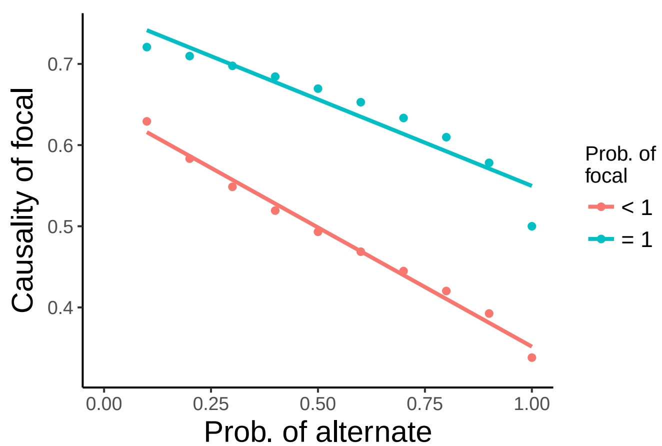

Figure S7: **(absence of) Effect b in the normalized model.**

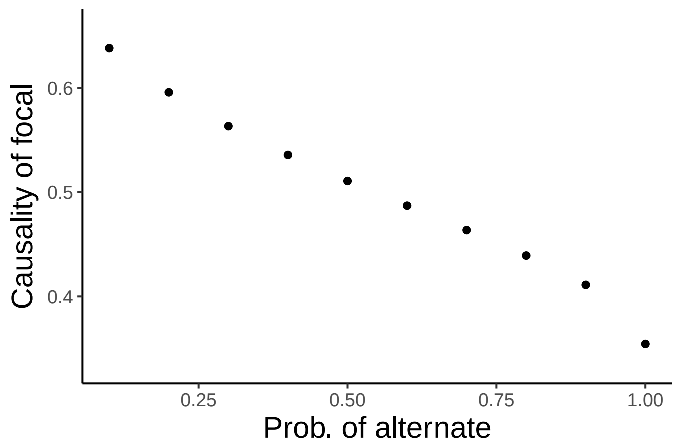

Figure S8: **Effect c in the normalized model.**

By contrast, consider the following normalization function, which does not rely on exponentiation, and therefore treats zeroes as 'true zeroes':

$$\tilde{k}'_{G\to D} = \frac{k_{G\to D}}{k_{G\to D} + k_{B\to D}}$$

When using this normalization function, our model exhibits all three non-linear effects described above (see figs. S9 to S11).

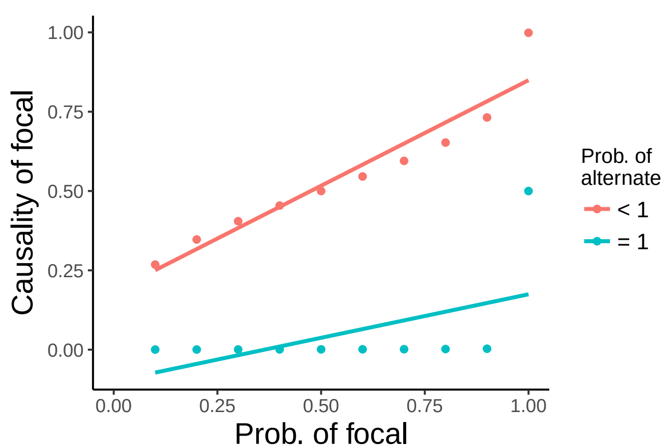

Figure S9: **Effect a in the model with alternative normalization.** Abnormal deflation ceases to occur when the alternate is certain.

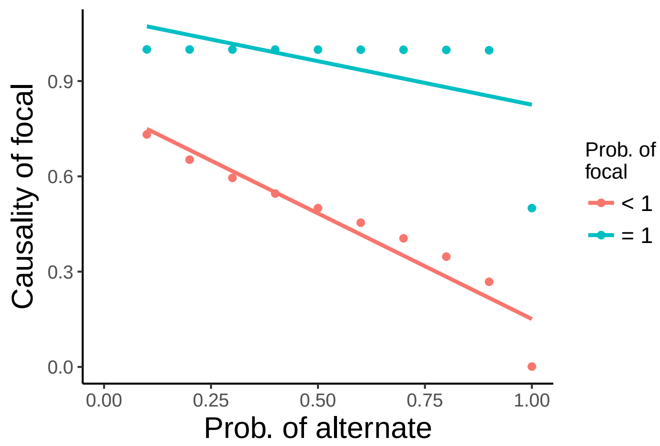

Figure S10: **Effect b in the model with alternative normalization.** Reverse supersession ceases to occur when the focal is certain.

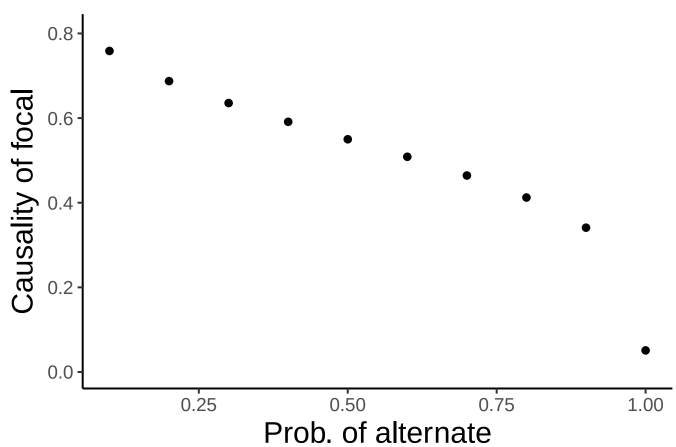

Figure S11: **Effect c in the model with alternative normalization.** Reverse supersession occurs mostly as the alternate variable approaches certainty (as $Pr(blue)$ approaches 1).

Applying this alternative normalization function to our model in the conjunctive structure does not change any of our main results (the model still shows the same fit to the data, and shows the same non-linear effects described in section 5.1; see Figure S12).

In the empirical data, people very rarely ascribe causal scores of 0, so they are probably not using the exact normalization function shown above. One may speculate that they use a normalization procedure which treats raw causal scores of zero as 'true zeroes' before comparing them to other raw causal scores, but then converts them into another number at the end of the procedure.

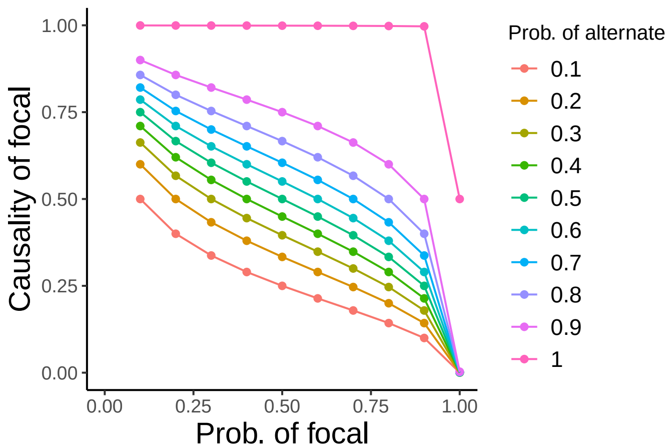

Figure S12: **Causality of green as a function of $Pr(green)$ and $Pr(blue)$, for model with alternative normalization, in the conjunctive structure.**

References

Halpern, J. (2016). *Actual causality.* MIT Press.

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

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions. *Psyarxiv*.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Springer.