Abstract. When judging what caused an event, people do not treat all factors equally – for instance, they will say that a forest fire was caused by a lit match, and not mention the oxygen in the air which helped fuel the fire. We develop a computational model formalizing the idea that causal judgment is designed to identify "portable" causes – causes that are likely to generalize across a variety of background circumstances. Under minimal assumptions, the model is surprisingly simple: a factor is regarded as a cause of an outcome to the extent that it is, across counterfactual worlds, correlated with that outcome. The model explains why causal judgment is influenced by the normality of candidate causes, and outperforms other known computational models when tested against an existing fine-grained dataset of human graded causal judgments (Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS one, 14*(8).).

keywords. causation | causal selection | computational modelling | counterfactuals

When multiple causes contribute to an event, we tend to discriminate among them: for instance, we tend to say that the forest fire was caused by the match lit by a careless camper, but we regard the presence of oxygen in the air as a mere 'enabling condition' or 'contributing factor'. This suggests that we implicitly rank the different causes of an event, as if we computed the 'actual causal strength' of each of them.

Here we propose a model of how the mind computes actual causal strength. Researchers have proposed that cognitive mechanisms for causal judgment are well-designed for the problem of identifying 'portable' causes, i.e. causes that would reliably lead to an outcome, across a wide range of different background conditions (see (Lombrozo, 2010; Hitchcock, 2012)). For instance, the lit match is a 'portable' cause of the forest fire, because across a variety of plausible background circumstances, striking a match inside a forest would result in a forest fire.

We formulate this hypothesis as a simple computational theory (Marr, 1982). Identifying portable causes requires that when one judges how much a factor C was causally responsible for an outcome E, one does not focus exclusively on what actually happened. One also needs to compute the effect that a manipulation of C would have had on E in a range of alternative possible situations. This suggests a measure of causal strength which is similar to the 'effect size' measures that scientists use in interpreting the results of an experiment. On average, across possible situations, by how many standard deviation units can one change the value of E by making a one standard-deviation change in C? In many contexts, this is simply equivalent to computing the correlation between C and E across the possible situations that we imagined.

We formally express this theory as a simple algorithm, and show that it can explain a wide range of human causal intuitions.

1 Model

We define an algorithm which takes as input an event (e.g. someone lits a match, there is oxygen in the air, and the forest catches fire), and delivers an actual causal strength score for a candidate cause (e.g. how well the lit match qualifies as having caused the forest fire). We assume that the agent making a causal judgment possesses a representation of the causal structure of the situation she is evaluating (e.g. she knows that lightning a match tends to generate fire, unless there is no oxygen in the air). We use the formalism of structural equation models to model such representations (see SI for an informal introduction, and (Halpern, 2016), for a technical treatment), and refer to a specific state of a causal system as a 'world'. 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 worlds by sampling the set of possible worlds, according to the prior probabilities of the exogenous variables (i.e., sample worlds in proportion to how likely each world is). For each such world, the values of the endogenous variables are then determined naturally 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 sampled worlds (for exogenous variables, this can simply be read off from the variable's associated probability distribution).

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.

The first step of the algorithm generates a large number of possible worlds, ensuring that we can look at the effect of C on E across a large number of different background circumstances, where these circumstances are represented in proportion to how likely they are to arise. The second step looks at each of these worlds in turn, asking about the strength of the causal dependence of E on C in each world. Our measure of causal dependence is standardized by the ratio of the standard deviation of C to the standard deviation of E. This standardization is akin to what scientists do when they compute statistical measures of effect size such as a Pearson's r; it allows measures of causal effects to be unit-free (so that, e.g., the causal strength of temperature does not depend on whether it is measured in Fahrenheit or Celsius). Finally, the last step of the algorithm takes the average of all the causal dependence scores computed in this way.

If C and E obey the "no-confounding assumption" (Pearl, 2000), then $k_{C\to E}$ is simply the correlation between C and E across worlds sampled in step **a** (we prove this for the case of binary variables in the SI). The "no-confounding assumption" holds when C has a causal influence on E, E does not have a causal influence on C, and no variable has a causal influence on both C and E.

Intuitively, when this assumption holds, the relationship between C and E is not confounded by third variables, so we can read the causal effect of C on E from the correlation between C and E even in 'observational' data (i.e. data which was generated without performing any intervention) (Pearl, 2000).

2 Comparison with human causal intuitions

When judging whether a factor is causal, people are sensitive to its statistical normality (i.e. its frequency, or its probability), as well as the statistical normality of other factors. The present model parsimoniously explains four qualitative effects of normality on human causal judgments, most of which have been replicated many times across different contexts. We show below that it also provides a good quantitative fit to fine-grained data from a recent set of experiments (Morris, Phillips, Gerstenberg, & Cushman, 2019). For reasons of space, we also describe the four qualitative effects in the context of the Morris et al. (2019) set of experiments, since these experiments exhibited all four effects.

Morris et al. asked participants to read the following vignette:

> A person, Joe, is playing a casino game where he reaches his hand into two boxes and blindly draws a ball from each box. He wins a dollar if and only if he gets a green ball from the left box and a blue ball from the right box. Joe closes his eyes, reaches in, and chooses a green ball from the first box and a blue ball from the second box. So Joe wins a dollar.

Participants were asked to rate, on a 1-9 scale, their agreement with the statement "Joe's first choice (where he chose a green ball from the first box) caused him to win the dollar".

In a first experiment, participants saw the vignette shown above, which describes a conjunctive structure (Joe needs to draw a green ball from the first box, AND a blue ball from the second box, in order to win). In a second experiment, another set of participants read the same vignette, minimally modified so as to depict a disjunctive structure (Joe needs to draw a green ball from the first box, OR a blue ball from the second box, in order to win).

Participants were shown pictures of the two boxes. Across conditions, the experimenters systematically varied the proportion of green balls in the first box and blue balls in the second box. The proportion of green balls in the first box varied from 0.1 to 1, in 0.1 increments; the proportion of blue balls in the second box was similarly and independently manipulated. Morris et al. (2018, 2019) assessed the fit of prominent existing computational models of causal judgment (Icard, Kominsky, & Knobe, 2017; Halpern & Hitchcock, 2015; Morris et al., 2018; Cheng, 1997; Jenkins & Ward, 1965; Spellman, 1997) to their dataset.

Following (Morris et al., 2018), we generated predictions for two versions of our model. The first version is the baseline version of the model. The second version is a "normalized" version, generated with the softmax function:

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

Where $k_{G \to D}$ is the baseline causal strength ascribed to the draw of the green ball, and $k_{B \to D}$ is the baseline causal strength ascribed to the draw of the blue ball (Morris et al., 2018). We also considered a baseline and a normalized version for all the models that are studied in Morris et al. (see Morris et al., 2019, 2018 for a description of these models). For each causal structure, we computed the predictions of our model by deriving analytical expressions corresponding to the correlation between "Joe draws a green ball" and "Joe wins a dollar" in the limit of an infinity of samples (see SI for derivation). We studied the performance of each model in each causal structure by computing the item-level correlation between a model's predictions and participants' average causal ratings.[1]

2.1 Conjunctive structure

Results are shown in Figure 1. Both the human data and the model exhibit two well-known effects of statistical normality on causal judgment. The first effect is *abnormal inflation*: as "drawing green" becomes less likely, causality ratings for "drawing green" increase (Hilton & Slugoski, 1986; Kahneman & Miller, 1986). The second effect is *supersession* : as"drawing blue" becomes more likely, causality ratings for "drawing green" increase (Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015).

[1]R code to reproduce analyses and figures is available in the electronic supplementary materials.

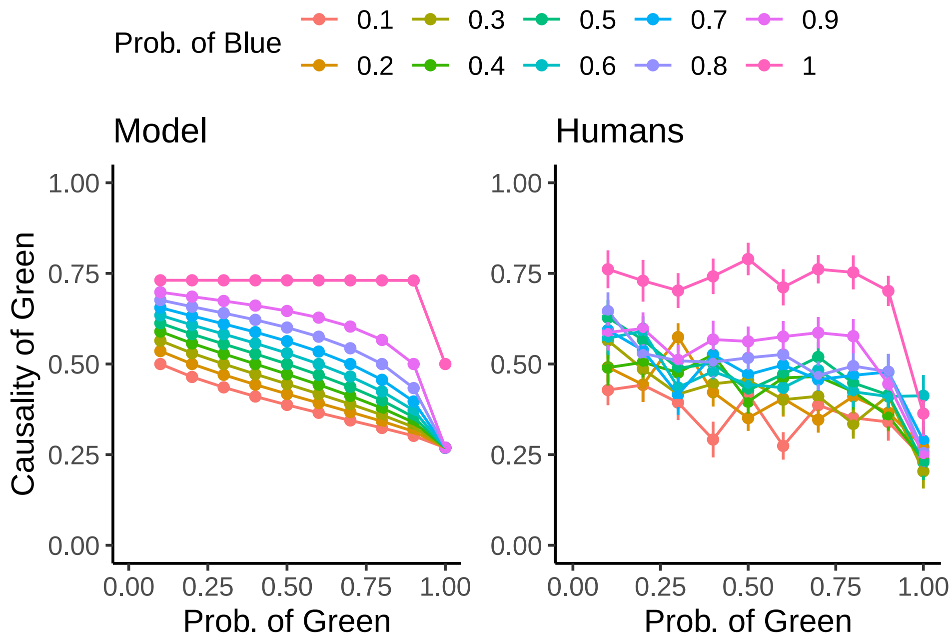

Figure 1: **Judgments made by the normalized version of the model in the conjunctive structure, along with average human judgments.** Human data are from Morris et al. (2019), and are standardized on the [0,1] interval.

Figure 2 shows the fit of each model to the data. The normalized version of our model had a marginally better fit than the baseline version (William's t-test, $t(97) = 1.88$, $p = .06$), and a better fit than all other models (all $ts > 6.11$, all $ps < .001$).

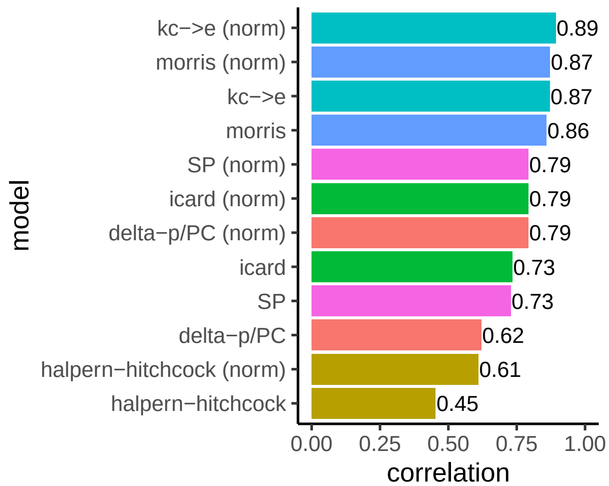

Figure 2: **Fit of each model to human data, conjunctive structure.**

2.2 Disjunctive structure

Results are shown in Figure 3. Both the human data and the model exhibit *abnormal deflation*: as “drawing green” becomes less likely, causality ratings for “drawing green” decrease (Icard et al., 2017; Gerstenberg & Icard, 2019; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019). They also exhibit an effect, *reverse supersession*, that had not been identified prior to the study by Morris et al.: as “drawing blue” becomes less likely, causality ratings for “drawing green” increase.

We note that the reverse supersession effect is relatively weak in the human data, and is mostly driven by cases where “drawing blue” is certain to occur; indeed, Kominsky et al. (2015), in a study with lower statistical power, and that did not include candidate causes that were certain to occur, were not able to find evidence for a reverse supersession effect. High-powered replications of the effect are a ripe area for future research.

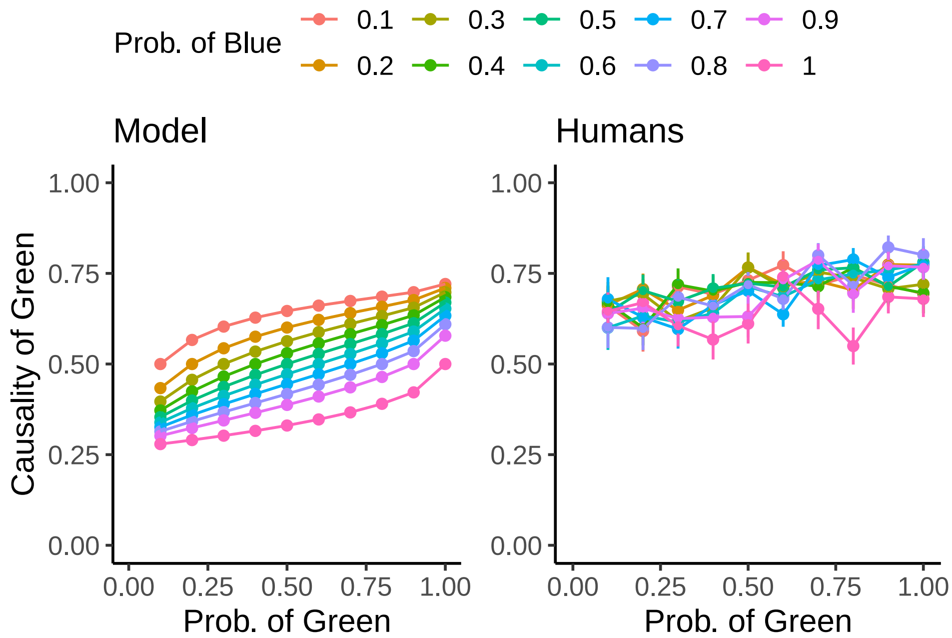

Figure 3: **Judgments made by the normalized version of the model in the disjunctive structure, along with average human judgments.** Human data are from Morris et al. (2019), and are standardized on the [0,1] interval.

Figure 4 shows the fit of each model to the data. The best performing models were the normalized version of our model, both versions of the Icard model and the normalized Delta-P model. None of these four models fit the data better than any other, all $ts < .73$, all $ps > .47$. The next best model was the baseline version of our model, which performed less well than the models above (all $|ts| > 3.69$, all $ps < .001$), but better than all other models (all $ts > 4.03$, all $ps < .001$).

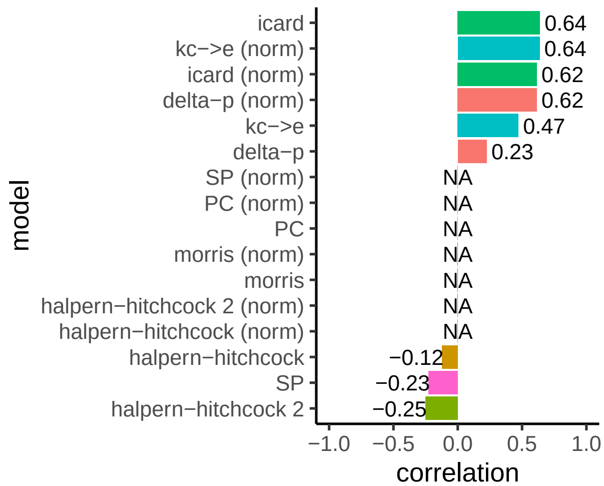

Figure 4: **Fit of each model to human data, conjunctive structure.**

Morris et al. (2019) also highlight interesting non-linear patterns in their data, for both experiments. Our model mostly reproduces these non-linear patterns (see SI).

3 Discussion

Our simple model provides a normative justification for the complex pattern of effects of statistical normality on causal judgment: causal cognition appears to be well-designed to identify 'portable' causes. Our work also provides a normative justification for the hypothesis that causal judgment relies on a process which samples counterfactuals according to their normality (Icard et al., 2017) [2].

Another recent measure of actual causal strength, the SAMPLE measure (Morris et al., 2018) can be easily derived from the present model. For

[2] at least as far as *statistical* normality is concerned; this could be extended to other types of normality using recent arguments by (Phillips, Morris, & Cushman, 2019)

any causal structure in which C and E are binary variables obeying the no-confounding assumption, and C is necessary for E, the SAMPLE measure is equivalent to the square of $k_{C \to E}$ (see SI).

Many existing measures of actual causal strength are based on the notions of necessity and sufficiency (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Icard et al., 2017; Morris et al., 2018). Necessity and sufficiency are not primitives in our model, but in the special case where we assume binary variables, then the $\frac{\Delta_E}{\Delta_C}$ term used by the algorithm reduces to a measure of sufficiency (when we consider an intervention setting C from 0 to 1) or a measure of necessity (for an intervention setting C from 1 to 0): C is sufficient (or necessary) for E if $\frac{\Delta_E}{\Delta_C} = 1$.

Why is causal judgment well-designed to identify portable causes? The present results are consistent with several possibilities. Morris et al. (2018) recently argued that causal judgment serves to identify our best-bet intervention if we want to bring about an outcome but do not know the exact state of the causal system. Our model is consistent with this argument. On average, we can expect that an intervention on C will result in a change of $k_{C \to E}$ standard deviation units in E for each one standard deviation unit change in C. Therefore, if we want to set E to a certain value, we are generally better off making an intervention on the variable X with the highest $k_{X \to E}$. However, identifying portable causes may also be useful for a broader range of cognitive activities, such as prediction or explanation. The proper evolutionary domain of causal judgment remains an open question.

Closer to Marr's algorithmic level of analysis (Marr, 1982), future research should take a closer look at which actual causal strength measure best approximates human judgments. Our model had the best overall fit to the Morris et al. (2019) dataset, but other models (notably Icard et al., 2017) also performed well. It will be important to extend this comparison to a wider range of experimental setups (see e.g. (Sytsma, 2019) for preliminary evidence that differences in study design may influence causal attributions).

Although very general, our model is not a full theory of causal judgment. Just as other models of actual causal strength, it is relatively insensitive to the specifics of what *actually* happened. Imagine that Suzy and Billy throw a rock at a bottle, but Suzy's rock gets there first. Against intuition, the present model assigns positive causal strength to "Billy's rock broke the bottle", because there are possible worlds where Billy's rock would have made a difference to whether the bottle breaks. Future work should integrate the present ideas with theories which can handle such cases (e.g. (Halpern & Pearl, 2005)).

References

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological review*, *104*(2), 367.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the cognitive science society.*

Gerstenberg, T., & Icard, T. (2019). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*.

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

Halpern, J., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413–457.

Halpern, J., & Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, *56*(4), 843–887.

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, *93*(1), 75.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological monographs: General and applied*, *79*(1), 1.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information* (Vol. 2). Inc., New York, NY.

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.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323.

Sytsma, J. (2019). The effects of single versus joint evaluations on causal attributions. *PhilSci Archive*.