Supplementary Materials

Modeling and leveraging intuitive theories to improve vaccine attitudes

Derek Powell, Kara Weisman, and Ellen M. Markman

This project represents a line of theoretical and empirical work seeking to understand misconceptions and belief revision in the context of people's intuitive theories. We focus our efforts on reducing vaccine skepticism by identifying people's intuitive theories surrounding vaccination decisions. Through this case study we hope to shed light on why such misconceptions persist and how educators and interventionists might address misconceptions more effectively across domains. We argue that correcting misconceptions requires presenting evidence in a way that is sensitive to and persuasive within the broader conceptual context in which those misconceptions reside (see Gershman, 2019; Weisman & Markman, 2017). Understanding the revision of any one specific belief (e.g., a misconception like *vaccines are dangerous*) demands a holistic understanding of the wider set of related beliefs that make up people's intuitive theories.

First, we describe in further detail the modeling process for developing a Bayesian network model of the intuitive theory surrounding vaccination decisions.

Then, we describe our empirical methods in detail. This begins with the qualitative research we conducted with vaccine skeptics to develop psychometrically-valid quantitative scales measuring a network of 14 beliefs related to vaccine skepticism. Then, we report the detailed methods for each of the empirical studies reported in the paper. In Study 1, we used Bayesian network structure learning algorithms to develop a cognitive model of the relationships among these beliefs in a large sample of people. This model can predict how beliefs correlate across people and how they would be revised in the face of evidence. In Study 2, we replicated the success of an existing intervention emphasizing the dangers of childhood diseases (Horne et al., 2015) and used our model to provide a more detailed understanding of *how and why* this intervention increases vaccine intentions. In Study 3, we drew inspiration from our model to create and validate the success of a novel intervention aimed at debunking concerns about toxic additives. Finally, in Study 4 we tested whether people's beliefs changed coherently following relevant real-world events, by examining beliefs before and after serious outbreaks of measles in the United States in early 2019.

Theory of intuitive theories

Quine argued that all of one's beliefs are intertwined in a "web of belief" (Quine, 1951; Quine & Ullian, 1978)—an apt metaphor for intuitive theories. Intuitive theories consist of relations between beliefs that specify how states of affairs causally, logically, or otherwise interconnect with one another. Specifically, we consider intuitive theories to be generative mental models that entail probabilistic dependencies among possible states of affairs.

Given this definition, we argue that any pair of beliefs related by a intuitive theory can be expected to exhibit several qualitative features:

1. When two beliefs are related by an intuitive theory, evidence affecting one of those beliefs will also affect the other in accordance with their relationship in the intuitive theory (e.g. evidence for B will increase credence in A, and vice versa). This is a clear and direct consequence of the laws of probability and Bayes' rule.

2. Where beliefs about two states of affairs, A and B, are related through an intuitive theory (e.g., A causes B), those beliefs will be systematically correlated across individuals in proportion to the average conditional probabilities (e.g. people who believe A will also tend to believe B);

In the case of beliefs about binary states of affairs, this can be seen through an application of the law of total probability and some algebra, which reveals a linear relationship between the two beliefs across individuals. To derive this result, we assume each individual i holds a specific degree of belief in x and y, $p_i(x)$ and $p_i(y)$. Moreover, these degrees of belief are determined by an intuitive theory: a generative mental model encoding a conditional dependency between the two, i.e. $p_i(y|x)$ and $p_i(y|\neg x)$.

From the law of total probability, we can write:

$$p_i(y) = p_i(y|x)p_i(x) + p_i(y|\neg x)(1 - p_i(x)) \tag{1}$$

With some algebra, we can rearrange these terms:

$$p_i(y) = p_i(y|\neg x) + (p_i(y|x)p_i(x) - p_i(y|\neg x))p_i(x) \tag{2}$$

This equation has a familiar form $y = \alpha + \beta x$. We know that a least-squares solution for β will return $\frac{1}{n}\sum_i p_i(y|x)p_i(x) - p_i(y|\neg x) = \overline{p_i(y|x) - p_i(y|\neg x)}$, and that such an estimate for $\beta \propto corr(p_i(x), p_i(y))$.

3. Thus we can often reasonably expect there will be some systematic relationship between the average change in each belief following evidence and the correlation among those beliefs across individuals. More specifically, if we assume that the changes in beliefs A following evidence are uncorrelated with the conditional probabilties connecting A and B in people's intuitive theories, then the average change in belief A and B following evidence affecting A will be proportional to the correlation of those beliefs across individuals.

Again we consider this in the case of beliefs about binary states of affairs. We define $p_i(x)$ to be the prior probability of x before observing some evidence, and $p_i(x|e)$ to be the posterior probability of x after observing the evidence, e.

We define the change in $p_i(x)$ as $\Delta p_i(x) = p_i(x|e) - p_i(x)$.

We can write $\Delta p_i(y)$ in terms of $\Delta p_i(x)$. We assume that the evidence has a direct impact on the probability of x, but that the evidence only affects $p_i(y)$ through its impact on $p_i(x)$, i.e. $y \perp e|x$. This implies that $p_i(y|e) = p(y|x)p(x|e)$.

First, we expand $\Delta p_i(y)$ via the law of total probability and rearrange terms

$$\begin{aligned}
\Delta p_i(y) &= p_i(y|e) - p_i(y) && (3)\\
&= p_i(y|x)p_i(x|e) + p_i(y|\neg x)(1 - p_i(x|e)) - p_i(y|x)p_i(x) + p_i(y|\neg x)(1 - p_i(x)) && (4)\\
&= (p_i(x|e) - p_i(x)) \cdot (p_i(y|x) - p_i(y|\neg x)) && (5)
\end{aligned}$$

Taking the average, we have

$$\overline{\Delta p_i(y)} = \overline{p(y|e) - p(y)} = \frac{1}{n}\sum_i (p_i(x|e) - p_i(x))(p_i(y|x) - p_i(y|\neg x)) \tag{6}$$

From the definition of covariance we have

$$cov(X, Y) = E(XY) - E(X)E(Y) \tag{7}$$

If we assume that the change in belief in x following the evidence, $\Delta p_i(x)$ is uncorrelated with the conditional probabilities encoded in the intuitive theory $p_i(y|x)p_i(x) - p_i(y|\neg x)$ (i.e. covariance equals zero) and apply this definition, we can then rewrite the average change as

$$\overline{\Delta p_i(y)} = \overline{p(y|e) - p(y)} = \frac{1}{n}\sum_i p_i(x|e) - p_i(x)\frac{1}{n}\sum_i p_i(y|x) - p_i(y|\neg x) \tag{8}$$

And we have thus shown that $\overline{\Delta p_i(y)} \propto \overline{p_i(y|x)p_i(x) - p_i(y|\neg x)} \propto corr(p_i(x), p_i(y))$

In practice we did not measure probabilistic credences directly, but rather asked participants to report on a agree/disagree scale. We assume these actual measurements are some monotonic function of these credences, $f(p_i(x))$. Allowing that f could be non-linear, the proportionality conclusions are no longer guaranteed, though a monotonic association would remain.

Modeling

Bayesian network parameterization

Figure 1 shows an example of a simple Bayesian Network. The network itself is a *directed acyclic graph* (DAG), a type of graphical model. Graphical models represent variables as nodes connected by edges. In a DAG, these edges are directed, flowing from parent to child, and are constrained so that they cannot form cycles or loops. In a Bayesian network, these directed edges encode conditional dependencies among variables, so that the conditional probability distribution for a given variable is defined only in terms of its direct parents. Together, a child node and its parents are called a *family*.

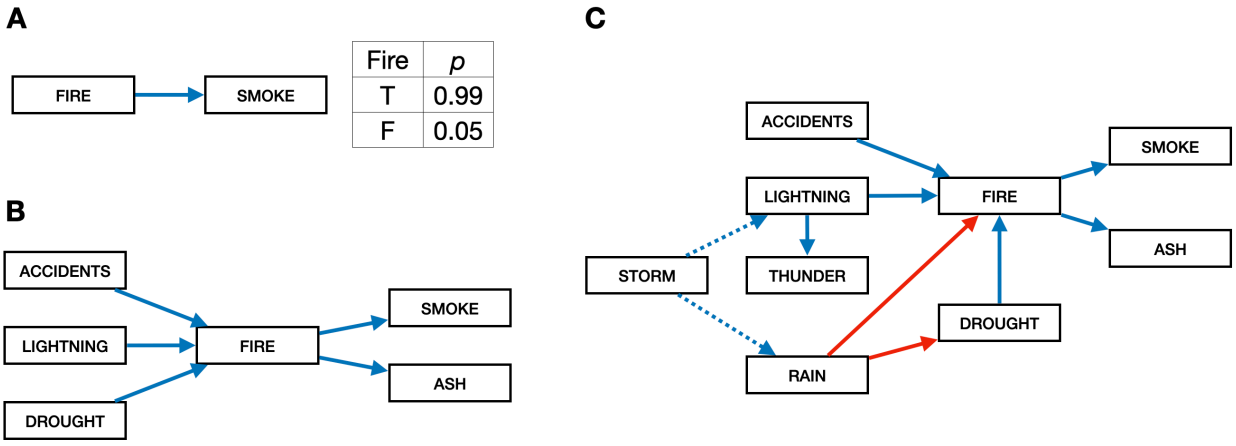

Figure 1: Several examples illustrating different features of Bayesian Networks.
A) A graphical representation of influence relations: Fire causes smoke. This influence is represented by the conditional dependencies captured in a conditional probability table (CPT).
B) A "parent" node can be linked to multiple "children" and a "child" node can be linked to multiple "parents": A child and its parents together compose a "family". This graph has three families: Accidents, lightning, drought, fire, fire, smoke, and fire, ash described by a corresponding CPT (not pictured).
C) Not all influence relations correspond to causation: A storm consists of rain and lightning (marked as dashed arrows). Causes can be generative or preventive, e.g. rain prevents drought and fire (red arrows). Some nodes may exert countervailing influences on others: A storm consists of rain, which can prevent fires, but a storm also consists of lightning, which can cause fires. In addition, parents may influence their children independently or interactively. Lightning and accidents could be seen as independent causes of fire—an accident may cause a fire whether or not there is lightning, and vice versa. However, drought conditions interact with both of these other parent nodes, making them more likely to produce a fire. In the present work, we will consider only the role of independent influence relations.

In this graph, the edge $FIRE \rightarrow SMOKE$ captures that fire causes smoke.[1] This relationship is represented as a probabilistic function, where the probability of smoke depends on the probability of fire. The probability of a node Y can be defined from its parents X in many different ways. For binary nodes, the most flexible specification is a conditional probability table (CPT) with 2^n entries specifying the probability of Y given the entire joint probability distribution over X, where n is the number of parents X. Given the exponential complexity of the CPT, simpler functions are often used to specify the probability of the child node, such as the noisy-or and noisy-and-not functions. A noisy-or function assumes each of the parents are independent generative forces, each with a probability of generating the effect. A noisy-and-not function assumes each of the parents are independent preventive forces, specifying the probability of the child node as the probability that it is generated *and not* prevented by any one of its possible preventers. We construct our cognitive model parameterization using a combination of these two functions, using a distribution sometimes referred to as a "DeMorgan gate" (Maaskant & Druzdzel, 2008). In our model, the probability of the child node Y can be expressed as:

$$P(y) = \Big(1 - (1 - w_0) \prod_i (1 - w_i)^{x_i}\Big) \prod_j (1 - w_j)^{x_j} \tag{9}$$

Here, each parent is either a generative (i) or preventive (j) influence, that generates or prevents the child node with probability weight w_i or w_j. The weight w_0 represents the "leak" probability, or the baseline probability that the child occurs without any other generating influences. The first term in brackets represents the probability that y is generated, or alternately, that not all of the possible generating forces fail to generate it. The second term represents the probability that it is not then prevented, or that all the preventing forces fail.

A cognitive model composed of these types of relations is capable of capturing a causal system of non-interactive generative and preventive causes. Although people's intuitive theories are likely more flexible than this, there is ample evidence that people are highly capable of reasoning according to these types of causal relationships (see Holyoak & Cheng, 2011 for a review), and moreover, that their causal learning is strongly biased toward learning these kinds of relations (e.g. Novick & Cheng, 2004).

This gives us a formalism for representing relations within an intuitive theory. Now we need methods for estimating the relationship that are actually represented within people's intuitive theories. Given measures of participants' credences in each of the parent beliefs x_i and the child belief y, we assume the following beta regression model:

$$y \sim beta(\mu k, (1 - \mu)k) \tag{10}$$

$$\mu = \exp\left(\ln\Big(1 - \exp\big(\ln(1 - w_0) + \sum_i (1 - w_i)x_i\big)\Big) + \ln\Big(\sum_i (1 - w_i)(1 - x_i)\Big)\right) \tag{11}$$

Via maximum likelihood estimation, we estimate for each parent a discrete parameter x_i determining whether the relationship is generative or preventive, and a continuous weight w_i determining its strength. We also estimate w_0, the "leak probability" which is assumed to be generative, and k, a measure of dispersion in the responses y.

One way to satisfy the assumption that changes in some belief following evidence are uncorrelated with the conditional probabilties connecting that belief to others in people's intuitive theories is to assume that all people hold the same conditional probabilities in their intuitive theory. This is essentially the assumption that we are making by assuming a single network structure where parents are parameterized as independent generative and preventive forces on their children. Thus, the resulting network we develop can be expected to make predictions in accord with point #3 above: the average predicted change in each belief following

[1] The example of cause and effect is especially intuitive, but these edges need not be causal. Instead, they are better thought of as representing "influence relations," which may also include relationships like logical implication and set membership (Williamson, 2001).

evidence directly impacting another belief will be proportional to the correlation between those beliefs across people.

Structure learning

Where the structure of an intuitive theory is known, Bayesian networks offer powerful tools for predicting and understanding how evidence will affect people's beliefs. However, there are many cases where the structure of the intuitive theory is unknown and is itself the very question under investigation.

Fortunately, the formalism of Bayesian Networks provides tools for navigating these challenges. Here, we used Bayesian network structure learning algorithms to construct a model of the intuitive theory of vaccine decisions (see Scutari & Denis, 2021). One way to learn a Bayesian network structure from data is to search through a space of networks, "score" each and choose the best scoring network. Many scoring metrics are possible, but a directly meaningful one is p(d|G), the likelihood of the data given the graph structure. Asymptotically, this can be estimated using the Bayesian Information Criterion (BIC):

$$BIC = k\ln(n) - 2\sum_i \hat{\ell}(x_i|x_{va(x_i)}) \tag{12}$$

Which we estimate using the maximum log-likelihood estimated for our statistical model, calculated as the sum of the maximum log-likelihood for each node given its parents, from a total of k parameters and n data points.

Generative model principle

We constrained our search by applying what we call the "generative model principle," assuming that the intuitive theory is a generative model, so that influence relations (edges) should flow from "generating" states of affairs toward states of affairs that result, such as from cause to effect.

We used "abstractness" as a heuristic for constraining structure learning according to the generative model assumption: We assume that abstract states of affairs generate more specific and concrete states of affairs, so that edges should flow from the abstract toward the concrete. For instance, an abstract belief, such as the belief that natural things are better than artificial things, could influence more concrete beliefs, such as the belief that vaccines are dangerous. We consider "naturalism" to be an abstract belief because it is a belief about two large classes of entities, "natural things" and "artificial things." The belief that vaccines are dangerous is relatively more concrete: the class of "vaccines" is relatively smaller and largely subsumed by the class of "artificial things." Here, what is true of a larger class influences what we should think about the more specific class.

Before constructing our model, we sorted the 14 measured beliefs into "tiers" based on how broad or abstract each belief was. For instance, we considered holistic balance and naturalism to be the most abstract beliefs among those we measured, and labeled these "worldviews"; we considered our outcome of interest, vaccine intentions, to be the most concrete measurement of a specific intended action. In between, we judged "medical skepticism," "parental protectiveness," and "parental expertise" to be more general than beliefs about diseases and vaccines, and so separated these into two tiers as well. We then applied the generative model principle to induce a partial-ordering of these beliefs, and a corresponding "blacklist" stipulating that certain edges must not appear in the final DAG. For instance, that there is no edge from vaccine intentions to naturalism. This constraint can be thought of as a strong prior over the possible structures of this intuitive theory (namely, $P(G) = 0$ for any graph G not respecting the principle).

Cross validation of structure learning results

To conduct an efficient search through the super-exponential space of DAGs, we tested several variants of two structure learning algorithms implemented in the "bnlearn" R package (Scutari, 2010). The first is

a purely score-based "hill-climbing" algorithm(HC) and the second was a hybrid algorithm that combined the hill-climbing algorithm with a prior "restriction" phase that pruned down the space of DAGs (MMHC). This pruning was accomplished with the "min-max parents and children" algorithm, which can be tuned with a hyperparameter alpha, which we tested at several levels. To speed computations, we restricted the maximum number of parents in the graph to five for all nodes, as the complexity of the scoring function is exponential in the number of parents. In all resulting networks no node was found to have more than four parents, suggesting this did not impact the final results.

We performed a series of 10-run 10-fold cross validation analyses to evaluate the success of these different structure learning strategies. In addition to the different algorithms, we also compared four different "blacklists" that imposed different constraints on the structure learning process. In addition to our theory-driven blacklist based on our generative model assumptions ("theory"), we also explored unconstrained structure learning, and two additional blacklists: one which stipulated only that "intentions to vaccinate" should be a child node with no children of its own ("child"), and another that stipulated also that the most abstract beliefs "holistic balance" and "naturalism" should be allowed to have no parents other than each other ("parents").

Figure 2 displays the results of these cross-validation tests. As seen in the figure, all structure learning approaches produced models with similar out-of-sample performance, with more variability among folds and runs than between algorithms. Thus the cross-validation does not uniquely identify an optimal structure learning strategy.

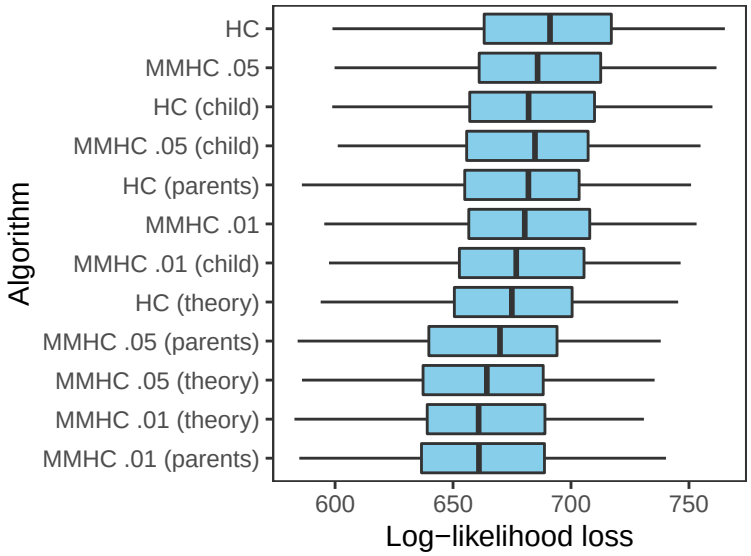

Figure 2: Cross-validation performance for diffrent structure learning algorithms on the training data from Study 1.

Nevertheless, there were substantial differences among the approaches, especially from the application of different constraints through blacklists. In particular, structure learning conducted without any constraints produced a very different model than structure learning under the generative model constraints; although many edges were shared, the directions of these edges were very frequently reversed. For instance, in structures learned without constraints, vaccine intentions is very frequently a parent node with many children, rather than a child of other nodes. These findings underscore the need for top-down constraints in the use of structure learning algorithms to identify plausible cognitive models for relations among beliefs. Models identified without these constraints may serve fine for prediction, but appear unlikely to lend themselves to clear interpretations as cognitive models.

Without any clear guidance from cross validation and with a clear need to apply top-down constraints, we elected to advance our modeling with a model learned with the hill-climbing algorithm and our generative

model constraints. This model contained more arcs than models estimated with hybrid algorithms, but because these hybrid algorithms include a restriction phase based on correlations rather than our statistical model, we preferred the unrestricted approach.

Interpreting the model

We used a bootstrapping procedure to estimate the confidence in each of the links in this model, conducting the structure learning procedure repeatedly for 200 bootstrapped resamples of the original training data. Figure 3 visualizes the frequency with which each arc appeared in the resulting model across these replicates, and Figure 4 visualizes the model with the coefficients for each link encoded by color.

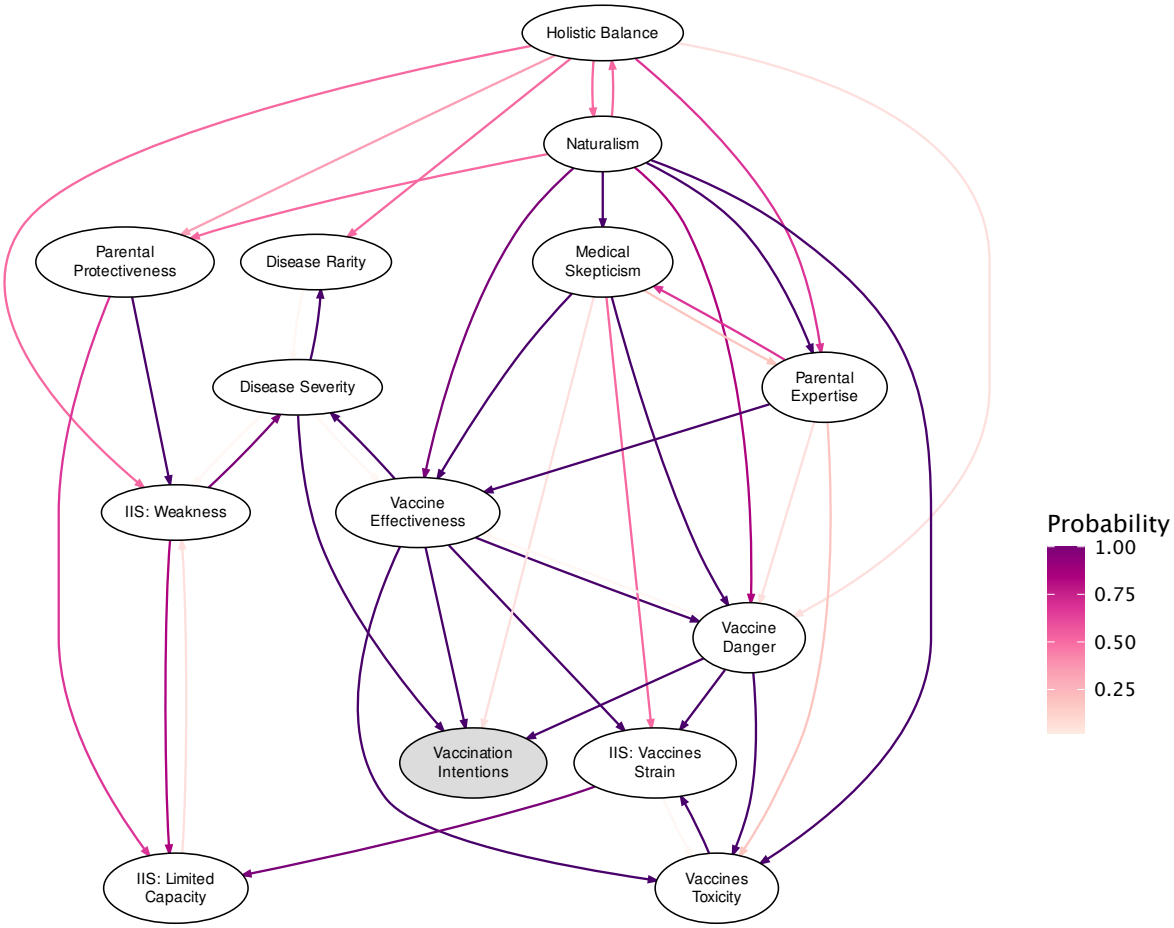

Figure 3: Bootstrapped confidence estimates for Bayesian network edges. Edge color indicates frequency with which edges appeared when conducting search across 200 bootstrap replicates. Only edges appearing in at least 20% of graphs are shown. Our primary outcome of interest, vaccination intentions, is highlighted in gray.

Though we found the uncovered network structure to be quite revealing and informative, there are also some important limitations that should be acknowledged. In particular, the edges from the vaccine danger and vaccine effectiveness nodes to the *vaccine toxicity* node strike us as counterintuitive. Intuitively, we might imagine that vaccines having toxins could cause them to be dangerous and ineffective. Yet the edges in the model run from in the opposite direction, and so do not support a clear causal interpretation.

A Bayesian network is a factorized representation of a probability distribution and any given network is a member of a "Markov equivalence class"—a set of networks capable of factorizing the same distribution. This set of DAGs can be represented as a "completed partially directed acyclic graph" (CPDAG), a graph with both directed and undirected edges. The equivalence class for our model can be represented as a CPDAG with 14 directed edges and 15 undirected edges—indicating that the data do not constrain the directions of 15 of the edges in the final model. However, 10 of these edges' directions are directly constrained by our top-down "generative model" constraints, which in turn constrains all but two of the remaining undirected edges, including the counterintuitive edges running toward *vaccine toxicity*. This is also reflected in a bootstrapping analysis applying our chosen structure learning algorithm against 1000 bootstrapped replicates of our data. We found that the direction of some edges within the graph differed across bootstrapped runs (Holistic balance → naturalism, medical skepticism → parental expertise), but that these relationships surrounding *vaccine toxicity* were robust.

Altogether, these findings might be seen as demonstrating the need for a heavier hand in establishing top-down constraints on the model fitting. Certainly, they call for caution in interpreting the relationships represented in this model. Due to their tight connections with structural causal networks, it is tempting to interpret the edges in Bayesian networks as "causal." However, these relationships are best thought of as probabilistic dependencies or "influence" relations (Williamson, 2001). Indeed, there are other cases where it is unclear if an edge should be directed at all: e.g., it is not obvious to us what would be the "correct" direction for the edge between naturalism and holistic balance. We might even speculate there is an additional unmeasured variable that connects these beliefs and that would best account for their association. Ultimately, this model should be understood as offering a plausible, but not definitive, account for a cognitive model connecting these different beliefs.

What does the resulting model reveal about the intuitive theory undergirding vaccination decisions?

The three nodes with direct connections to *vaccine intentions* are beliefs about *vaccine effectiveness*, *vaccine danger*, and the severity of childhood diseases (*disease severity*)—concrete beliefs that are, at face value, especially closely related to vaccination decisions. The directionality of these edges is informed by the constraints we applied via the generative model principle, but the identification of these three factors as the most proximal to vaccine intentions is a product of the structure learning approach. This gives us some confidence that the model is capturing important relationships.

Other parts of the model shed new light on the role of lay theories in vaccine decisions. For example, *naturalism*——the general view that natural things are better than artificial things—appears to be strongly related to *medical skepticism* and *parental expertise*; all three of these abstract beliefs are related to concrete beliefs that, in turn, feed into participants' vaccination intentions. This hierarchical structure is at least partly a product of our constraints that sorted these beliefs into "tiers" of abstractness. However, the fact that a model with these constraints still functions well for prediction demonstrates that this hierarchical structure is plausible.

Finally, this model structure also helps explain some initially surprising findings. For example, before collecting these data we speculated that dispelling the misconceptions about the capacity of the infant immune system could promote positive attitudes toward vaccination. We were disappointed to observe the weak first-order correlation between this belief and vaccine intentions in our behavioral data (r = -.09 in our training split). The model sheds light on this surprising (lack of) relationship: Although the belief that the infant immune system is limited in capacity is positively related to the belief that vaccines strain the immune system—discouraging vaccination, as we had assumed—it also seems to promote the belief that childhood diseases have severe consequences for young children (*disease severity*), which might, in turn, encourage vaccination. These countervailing forces—which, on average, seem to have canceled each other out in this sample—dissuaded us from our plans to develop an intervention aimed at dispelling misconceptions about limited capacity in order to encourage childhood vaccination.

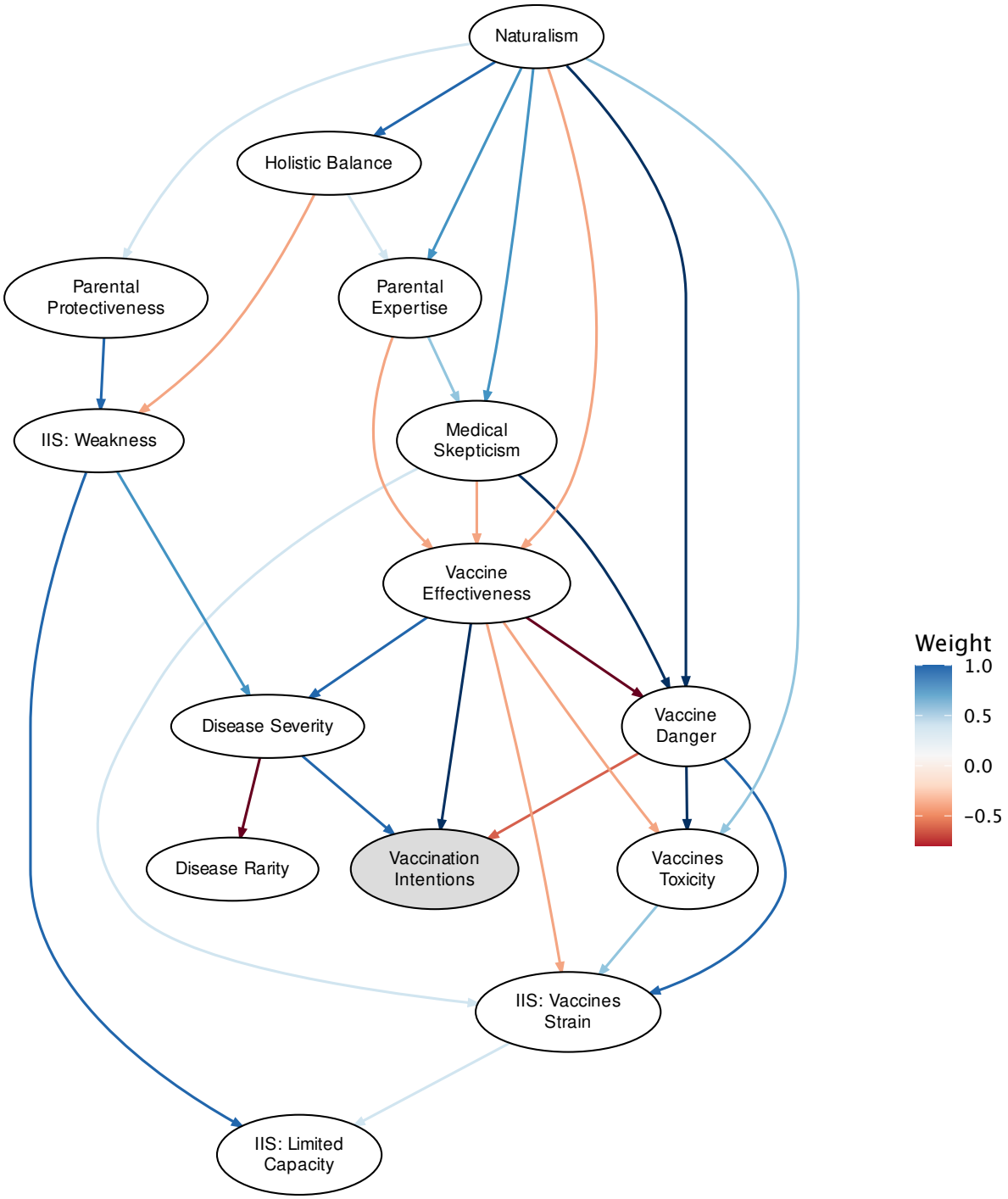

Figure 4: Bayesian network model of intuitive theory surrounding vaccination decisions. Edge color indicates the magnitude of the generative (blue) or preventive (red) coeffient for the links. Our primary outcome of interest, vaccination intentions, is highlighted in gray.

Altogether, these findings might be seen as demonstrating the need for a heavier hand in establishing top-down constraints on the model fitting. Certainly, they call for caution in interpreting the relationships represented in this model. Due to their tight connections with structural causal networks, it is tempting to interpret the edges in Bayesian networks as "causal." However, these relationships are best thought of

simply as probabilistic dependencies or "influence" relations (Williamson, 2001). Indeed, there are other cases where it is unclear if an edge should be directed at all: e.g., it is not obvious to us what would be the "correct" direction for the edge between naturalism and holistic balance. We might even speculate there is an additional unmeasured variable that connects these beliefs and that would best account for their association. Ultimately, this model should be understood as offering a plausible, but not definitive, account for a cognitive model connecting these different beliefs.

Validating the model

We validated our final model by testing its predictions on the held-out testing data. The models' predictions are plotted against the true values for the testing data in figure 5 below. As seen in the figure, the quality of predictions varied across beliefs, but in many cases was quite strong.

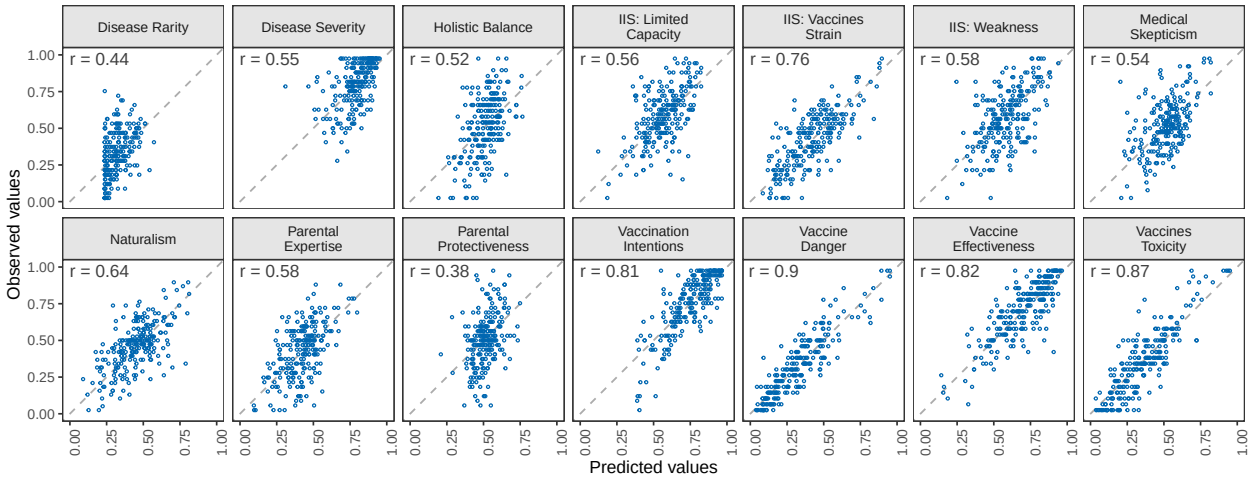

Figure 5: Model predictions versus observed values for test data from Study 1. Dashed lines represent perfect fits.

For each node, we also compared the Bayesian network model's predictions to a naive "maximal" model using values on all other belief scales as predictors. Our intention was to capture the degree to which each belief is predictable from the others as a whole. As shown in table 1, the relatively simpler Bayesian network model does not sacrifice any predictive power compared to the suite of maximal naive models.

Finally, we can also compare the quality of our model predictions to the test-retest reliability of the belief scales (estimated from the pre-test and post-test measures in the control condition of Study 2), establishing an overall upper-bound on the predictability of these measures. As shown in Table 1, the Bayesian network model did approximately as well as possible for seven beliefs, but fell short for the other seven. These failures suggest that the model is missing the information that would be needed to predict these beliefs. Although the notion of an intuitive theory presupposes that there is some collective unit of cognitive machinery defining people's thinking about a particular domain, defining the boundaries of an intuitive theory is far from straightforward. It could be that new measurements are needed to fully capture these aspects of the intuitive theory, or these beliefs are sufficiently peripheral that they should not be considered part of the intuitive theory. Given the fuzzy nature of the problem, and of cognition more generally, we expect that these types of decisions will be driven as much by investigators' pragmatic concerns as they are by psychological facts.

Table 1: Correlations for observed and predicted out-of-sample values

Belief scale	Bayesian network	maximal	test-retest
Vaccine Danger	0.898	0.895	0.928
Vaccines Toxicity	0.867	0.861	0.897
Vaccination Intentions	0.813	0.853	0.881
Parental Expertise	0.585	0.591	0.862
Vaccine Effectiveness	0.817	0.830	0.860
Naturalism	0.637	0.659	0.855
Medical Skepticism	0.544	0.617	0.842
IIS: Vaccines Strain	0.759	0.762	0.839
Holistic Balance	0.516	0.560	0.819
Disease Severity	0.550	0.562	0.784
IIS: Weakness	0.576	0.573	0.721
Disease Rarity	0.439	0.397	0.697
Parental Protectiveness	0.381	0.442	0.683
IIS: Limited Capacity	0.563	0.571	0.657

Estimating the strength of evidence from changes in belief reports

Returning to the DAG shown in Figure 1 above, consider the subgraph $FIRE \rightarrow SMOKE$, capturing the fact that fire causes smoke. This relationship is represented as a probabilistic function, where the probability of SMOKE depends on the probability of FIRE. If we observe the presence of SMOKE, we can draw inferences about FIRE.

However, sometimes reasoners have only indirect evidence at their disposal. For instance, a person might not directly observe smoke, but instead hear from a witness who claimed to have seen it. Indirect or partial observations can be captured by augmenting the network, creating the graph $FIRE \rightarrow SMOKE \rightarrow WITNESS$. Here, a new node WITNESS represents another person's testimony that they saw smoke, providing imperfect evidence for SMOKE, encoded by another probabilistic function. Observing WITNESS informs beliefs about SMOKE, in turn informing beliefs about FIRE. This approach to representing uncertain evidence is known as "virtual evidence" (Pearl, 1988).

To model the effects of participants observing evidence, we augmented the Bayesian network representing the cognitive model (shown in Figure 4) with an additional node representing the evidence. For instance, to model the effects of the "disease risk" intervention (Study 2), we added an evidence node as a child of "disease severity," acting as virtual evidence for this node. To compute the CPT for this node as a function of disease severity beliefs, we estimated an *evidence ratio* based on participants' pretest and posttest credences in disease severity. To so do, we made use of the log-odds expression of Bayes rule,

$$log(O(h|d)) = \ln\left(O(h)\right) + \ln\left(\frac{P(d|h=1)}{P(d|h=0)}\right) \tag{13}$$

Where the final term is the evidence ratio, defined as:

$$ER = \frac{P(d|h=1)}{P(d|h=0)} \tag{14}$$

The evidence ratio can be estimated from a model that is linear in the log-odds. This allows for estimation of the evidence ratio from the with the following beta-regression model:

$$y \sim beta(\mu k, (1-\mu)k) \tag{15}$$
$$logit(\mu) = logit(pretest) + \ln(ER_0) + x\ln(ER_1) \tag{16}$$

Where x was a binary variable representing the presence (1) or absence (0) of evidence at posttest, and participants' prior and posterior (pretest and posttest) credences were transformed by the logit function. This allowed us to estimate the log evidence ratio implied by participants' reactions to the intervention (ER_1) as well as the intervening time period without any explicit evidence (ER_0). Then, we constructed a CPT consistent with ER_1. For instance, with an evidence ratio below 1, we set:

$$p(evidence = 1|x) = .50 * ER_1 \tag{17}$$
$$p(evidence = 0|x) = .50. \tag{18}$$

We used this approach to predict how evidence (either interventions or real-world events) would change participants' beliefs in Studies 2-4 (see results for each study below). It should be noted that, with the exception of the data used to estimate this evidence ratio, these are out-of-sample and out-of-task predictions.

Methods

Qualitative Study

Before discovering a network connecting beliefs, we first needed to identify a set of beliefs that might plausibly influence vaccination decisions and develop ways of measuring these beliefs.

To do this, we drew on a variety of sources, including a qualitative survey with a set of pre-screened participants who had self-identified as vaccine skeptics. Out of 23 people invited to participate, n=16 participants completed the survey via Amazon Mechanical Turk (MTurk). All participants had previously answered "yes" to the question "Are you concerned that childhood vaccinations might cause serious side-effects?" When asked, "Do you have concerns about all vaccines or only some specific vaccines?" nearly all participants (14 of 16) expressed being concerned about specific vaccines; 3 of these participants, as well as the remaining 2 participants, expressed concern about all vaccines.

The remainder of the survey was devoted to open-ended questions, which we will summarize here as they were a major source of inspiration for the development of the 14 "belief scales" employed in Studies 1-4.

When asked about the potential side effects of vaccines, most participants (10 of 16) mentioned autism as a potential side effect; some mentioned other serious medical issues (e.g., stunting, cancer; 8 out 16); temporary, less serious physical symptoms (e.g., fever, headache; 7 of 16); other cognitive impairments and developmental disorders (e.g., learning disabilities, loss of speech; 6 of 16), and risk of death (6 of 16). Interestingly, 2 participants (of 16) specifically discussed the possibility that vaccines might have negative impacts on a child's immune system; for example, one wrote that a vaccine "penetrates and actually weakens the immune system overall (despite increasing immunity to specific sicknesses)." In explaining *how* vaccines might cause these side effects, many participants (9 of 15; 1 participant did not provide explanations) expressed concern that children's bodies and immune systems may be too weak or underdeveloped to tolerate vaccines (e.g., "The vaccines are given to the children when they are babies and their brains have not begun development"; "Too many are given all at once at too young an age"; "They overwhelm the child's nervous system"). Several participants (5 of 15) mentioned the presence of specific toxic additives in vaccines (e.g., "high levels of heavy metal, especially mercury"), or invoked the idea that vaccines are generally unnatural (e.g., "Vaccines are unnatural additions to the body. The human body should naturally fight off the disease and we should not be controlling how nature plays its role on mankind"; "These vaccines are simply chemicals that should not play a role in a young naturally developing child"). Several participants (4 of 15) expressed

concern that individual differences across children might render some children particularly vulnerable to these side effects (e.g., "The vaccine itself could interact with the specific child in a way that is unique and unforeseen"; "Sorry, I don't know the EXACT cause of it. I just think it has to do with the child's body being able to take the vaccine or not"). Individual participants also related personal stories of neighbors and friends, mentioned their mistrust of the government, or mentioned that the side effects of vaccines are worse than the symptoms of the diseases they are intended to prevent.

When asked whether vaccines were effective in preventing the diseases they were designed to prevent, participants were evenly split (8 said "yes" and 8 said "no") – but in both groups, participants' open-ended explanations touched on a common set of caveats and concerns. For example, some participants (7 of 15; 1 participant did not provide explanations) indicated that some vaccines are effective, at least for some individuals, but other vaccines are not (e.g., "Some work for some people, some don't"). Several participants (5 of 15) indicated that, regardless of efficacy, vaccines are no longer necessary because the diseases they prevent are no longer prevalent (e.g., "In certain circumstances they are, but in many cases the disease isn't really much of a risk so it doesn't make sense to fill the body with chemicals"; "I think that the potential of getting a disease is overblown and people get scared so they needlessly get vaccines"). Two (of 15) of participants mentioned that vaccines would be effective if they were given to children on a modified schedule (e.g., "Yes, if you give them ONE AT A TIME, spaced out over time and starting at older ages"); and two (of 15) related their personal mistrust of pharmaceutical companies (e.g., "THEY SEEMED TO BE A BIG MONEY MAKING SCAM BY THE BIG PHARMS"). (Individual participants related additional personal theories and concerns.)

When asked whether it would be safer to vaccine a 2-month-old child for 5 separate diseases in one doctor's visit (as the CDC recommends) or to spread these vaccinations out at a rate of 1 every 2 months, nearly all participants (15 of 16) said it would be safer to spread the vaccinations out. Many of these participants (8 of 15) indicated that this would be safer because it would prevent the immature system from being overwhelmed (e.g., "It gives a [child's] body time to adapt and doesn't overwhelm their tiny immune systems"); and several (4 of 15) suggested that this would be preferable because it would allow the parent more control over the process (e.g., "Ideally I would choose neither for my child, but if forced to choose between the two I would pick the latter. This would allow me to stop the vaccination process if things are not progressing well"; "I think a parent should have a choice which ones to have given to their child"). At least two participants expressed being deeply appalled by this choice, e.g., "He is only 8 weeks old. Who in their right mind decided it was safe to put 5 disease substances in a baby? Dead or alive, still horrific"; Honestly I do not think either is safe I think in both instances the child's body and health is actually being raped."

Finally, when asked why most healthcare providers recommend vaccinating young children, most participants (7 of 15; 1 participant did not respond) stated explicitly that healthcare providers have financial incentives to recommend vaccination – but many (7 of 15, including several who mentioned financial incentives) also acknowledged that healthcare providers likely believe (in the participants' view, mistakenly) that vaccinations are harmless and effective. Some participants (6 of 15) stated that healthcare providers are following protocol.

Drawing on these responses as well as anti-vaccine websites and academic articles on anti-vaccine skepticism Williams (2014), we generated a list of 13 underlying beliefs that might influence people's decisions about whether to vaccinate their children (*vaccine intentions*, the last of our 14 beliefs). These included a variety of claims about vaccines, including beliefs about (1) the overall danger of vaccines (*vaccine danger*); (2) *toxic additives in vaccines*; and (3) *vaccine effectiveness*, how effective vaccines are in preventing disease; as well as a variety of specific claims about childhood diseases like measles, mumps, and rubella, including beliefs about (4) *disease rarity* and (5) *disease severity*. Inspired by the many participants who invoked concerns about the effect of vaccines on infants' immature immune systems, we also developed scales to assess beliefs that (6) the infant immune system is weak (*IIS: weakness*); (7) the infant immune system is limited in its capacity and can be easily overwhelmed (*IIS: limited capacity*); and (8) vaccines strain the infant immune system (*IIS: vaccines strain*). Beyond this, we developed two scales to assess general theories about parenting, including (9) general *parental protectiveness*; (10) *parental expertise*, namely the belief that parents usually know more about their children's health than medical experts; and one scale to assess (12) *medical skepticism*, including concerns about pharmaceutical companies and corruption in the medical community. Finally, we decided to assess two broad worldviews: (12) *naturalism*, a general preference for natural over artificial things; and (13)

holistic balance, one important aspect of attitudes toward alternative medicine (e.g., "The body is essentially self-healing and the task of a health care provider is to assist the healing process"; (McFadden et al., 2010).

Holistic balance was the only belief for which we could identify a pre-existing psychometrically valid scale (McFadden et al., 2010). For each of the other 13 beliefs, we translated our qualitative insights into new psychometrically-valid scales through an iterative series of pilot studies not described here. We stipulated that each scale should be brief, composed of 4-6 statements for participants to evaluate; include at least one reverse-coded item; and be highly reliable in these pilot studies (observed reliability in final pilot study: all Cronbach's $\alpha \geq .80$).

General Methods: Studies 1-4

Studies 1-4 were conducted using online surveys (administered with Qualtrics survey software) with participants recruited from Amazon's mechanical turk (mTurk) work distribution website. All studies were conducted between August 2017 and June 2019. Repeat participation was prevented, with no participants involved in more than one study, except as noted in Study 4. A total of 2648 participants completed their participation in these studies. All participants had gained approval for $\geq$ 95% of previous work ($\geq$ 100 assignments); had verified US MTurk accounts; and indicated that they were $\geq$ 18 years old. See the participant demographics summary section below for full demographic information.

Method

Participants in all of these studies were asked to respond at least once to 14 belief scales measuring beliefs relevant to vaccination decisions (as previously described). Across all studies, these scales were each presented on a separate survey page. On each page, participants were asked to rate "how much you agree or disagree" on a 7-point scale from Strongly disagree (coded as -3) to Strongly agree (+3). Scales, and questions within each scale, were presented in a random order for each participant. Two to four attention checks ("Please select somewhat agree" and "Please select somewhat disagree") were embedded within two to four randomly chosen scales.

Data processing

For each scale, we generated a score for each participant by averaging their responses to each question in that scale (after reverse-coding where appropriate), and then rescaling these values to fall between 0 and 1, to represent a credence.

Study 1

Participants

A total of 1202 people recruited from Amazon Mechanical Turk (mTurk) participated in this study conducted in August of 2017. Participants were paid $1.60 for about 8 minutes of their time. 73 participants (6%) failed at least one of two attention check questions and were excluded from further analysis, leaving a final sample of n = 1129. After completing all 14 scales, participants were asked standard demographics questions as well as a handful of questions about whether they were or were expecting to be a parent. They were also debriefed with a brief statement noting that vaccines are safe and pointing them toward a WHO website for further information.

Results

Figure 6 presents correlations among the 14 belief scales measured in Study 1. As the figure shows, there are strong correlations among many of the beliefs, suggesting they are related by an underlying intuitive theory.

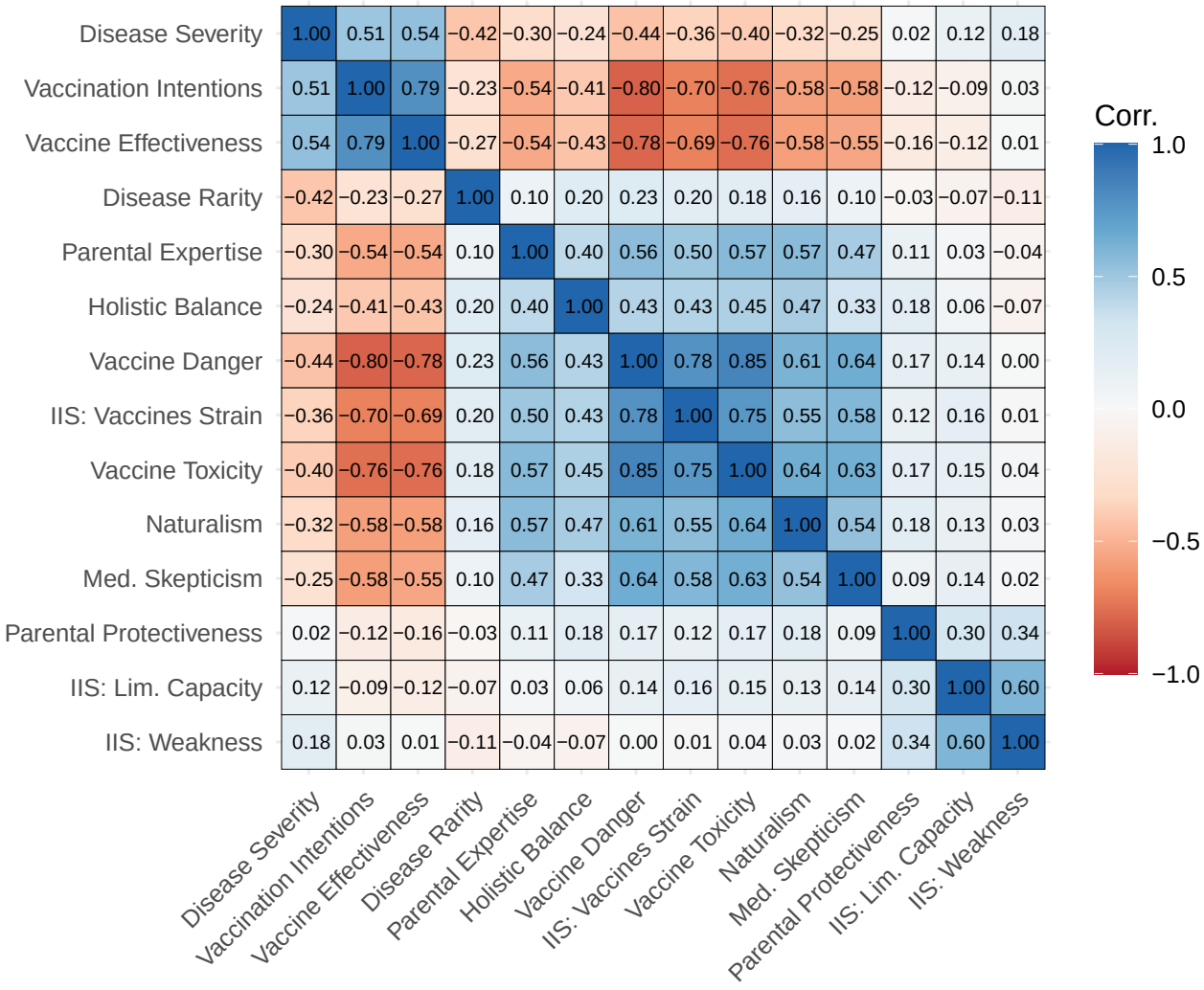

Figure 6: Correlations among 14 the belief scales measured in Study 1. Correlation matrix is arranged by hierarhical clustering algorithm to highlight connections among beliefs.

Study 2

Participants.

We recruited 860 people to participate in a two-part study via mTurk in April 2018. Two days later, participants who completed Phase 1 and passed attention checks were invited via email to participate in Phase 2 of the study, with 559 participants returning (726 invited, retaining 77% of the Phase 1 sample). Participants were paid $1.35 for their participation in each phase of the study. 54 participants (9.7%) failed at least one attention check in Phase 2 and were excluded from further analysis. We also excluded from analyses 12 additional participants who accessed the second survey more than once. This left a final sample of 493 participants.

Table 2: Bayesian regression model coefficients for model of Vaccine Intentions in Experiment 2.

Term	Estimate	$CI_{2.5\%}$	$CI_{97.5\%}$
α_1	-4.95	-6.49	-2.79
α_2	-3.70	-5.27	-1.53
α_3	-2.72	-4.29	-0.58
α_4	-1.59	-3.16	0.53
α_5	-0.37	-1.94	1.79
α_6	1.72	0.15	3.86
β_{phase}	-0.05	-0.23	0.12
$\beta_{condition}$	-0.07	-0.62	0.48
$\beta_{phase:condition}$	0.36	0.12	0.59
σ_{item}	2.14	0.93	4.58
σ_{subj}	2.88	2.66	3.11

Method.

Procedures for Phase 1 of Study 2 were nearly identical to Study 1, save for the use of four rather than two attention check questions, and the removal of debriefing statements at the end of the survey. In Phase 2 (approximately two days later), participants completed a study that was designed to be very similar to Horne, Powell, et al.'s (2015) original study. Participants were randomly assigned to either the disease risk condition or the no-intervention condition, with the constraint that the distributions of participants' Phase 1 *vaccine intentions* should be similar across the two conditions; we accomplished this by stratifying participants into groups based on their pretest vaccine intentions, and randomly assigning them to conditions in equal numbers within each group.

Participants were presented with a slightly modified version of the "disease risk" intervention from Horne et al.'s (2015) original study, which was based on materials available on the CDC website. Compared with Horne et al. (2015), we modified the intervention to present these parts in a fixed (rather than randomized) order, made slight changes to wording to improve clarity, and added a final statement summarizing the take-home message of the intervention. Full text of this intervention can be found in appendix B of these supplementary materials.

After reading this intervention, participants completed the same 14 belief scales that they had completed in Phase 1. Finally, participants were asked about their own vaccination behavior, including whether they had received a flu shot this season, and whether they planned to get a flu shot next season. They were also asked whether they were parents, how many children they had, and their children's ages. Parents were then asked additional questions about their vaccination decisions regarding their children, including whether their children had received flu shots and would receive them next season; whether they had ever chosen to delay a vaccine for their child; refused a vaccine for their child; or obtained an exemption for their child to attend school without a vaccination. At the end of the survey, participants were asked whether they had paid attention, avoided distractions, and taken the survey seriously.

Participants in the no-intervention condition ($n = 233$) completed an identical set of questions, except that they did not read any material before completing the 14 belief scales and the questions about their personal background.

Results

We conducted a mixed-effects Bayesian ordinal regression to assess the effects of the disease risk intervention on responses to the *Vaccination Intentions* scale items, using the "brms" package for R (Bürkner, 2017). We

regressed responses on phase (Phase 1 vs. 2, dummy-coded with Phase 1 as the baseline), condition (dummy-coded with the disease risk condition as the baseline, so as to assess difference scores in our condition of primary interest), and the interaction between phase and condition. We included random intercepts by subject and scale item. In this model, the unique effect of the intervention on vaccination intention responses is assessed by the interaction term (indicating the unique effect of Phase 2 for intervention versus no-intervention participants). The disease risk intervention increased participants' vaccination intention ratings (see Table 2). We consider this to be a full replication of Horne et al.'s (2015) results.

To model the changes in the belief network, we estimate the "evidence ratio" using the method described in the modeling section above. To simulate the impacts of this evidence, we supplied virtual evidence to the *disease severity* node by adding a child node to the cognitive-model network with a CPT capturing this evidence ratio. Given that our cognitive model is parameterized via noisy-logical functions, the predicted influence of each belief is closely related to its first-order correlation.

Figure 7 (right) shows the correlation between predicted and observed belief changes conditional on the disease risk intervention.

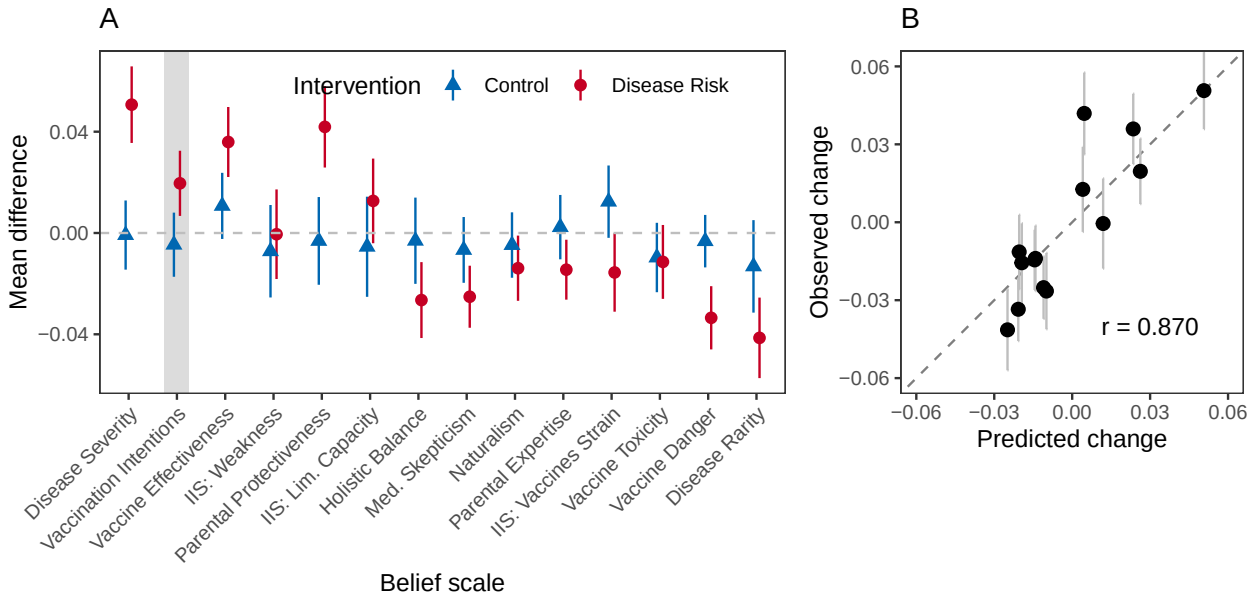

Figure 7: Belief changes following the intervention in Experiment 2. Left, mean difference scores from pretest to posttest for the 14 belief scales across experimental conditions. Belief scales are ordered by model-predicted changes based on the observed effect on disease severity. Our primary outcome of interest, vaccination intentions, is highlighted in gray. Right, mean observed and model predicted change scores for the disease risk intervention condition. In both plots, error bars represent 95% confidence intervals.

The models' predictions are strongly correlated with observed changes. Thus the model developed in Study 1 generalizes both across samples and across tasks: The model was developed from the correlations among beliefs in one sample of participants at a single time-point, and was able to accurately predict how a separate sample of participants revised their beliefs following evidence.

Studies 3a and 3b

In our paper, we laid out three qualitative features of intuitive theories.

1. Where beliefs about two states of affairs, A and B, are related through a shared intuitive theory (e.g.,

A causes B), those beliefs will be systematically correlated across individuals (e.g., people who believe A will also tend to believe B);

2. When two beliefs are related by an intuitive theory, evidence affecting one of those beliefs will also affect the other in accordance with their relationship in the intuitive theory (e.g., evidence for B will increase credence in A, and vice versa); and

3. Following from the prior points, across individuals, the average change in one belief following evidence affecting another belief will be proportional to the correlation between those beliefs across individuals (evidence for B will increase credence in A in proportion to the correlation between beliefs A and B).

By assuming that these beliefs are connected via an intuitive theory, we can infer how any given belief in the network will change in response to evidence targeting any other belief in the network. We used the cognitive model to predict how vaccination intentions would be influenced by an intervention targeting each of the 13 other beliefs in the network. Given that our cognitive model is parameterized via noisy-logical functions, the predicted influence of each belief is closely related to its first-order correlation. This captures qualitative Feature 3 of intuitive theories. Figure 8 presents these predictions, assuming the changes in beliefs are designed to move vaccine intentions in a positive direction (e.g. increasing "disease severity" beliefs but decreasing "vaccine danger" beliefs).

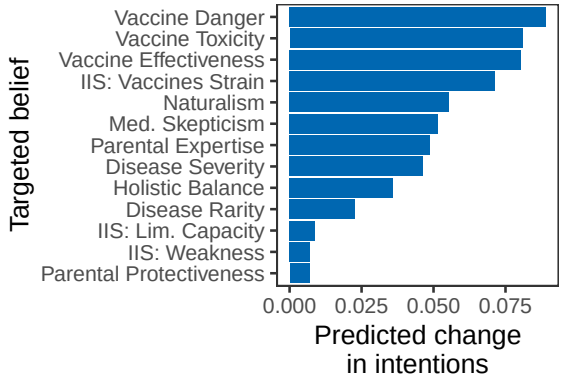

Figure 8: Predicted effects on Vaccine Intentions for hypothetical interventions targeting each of the 13 other beliefs

Guided by these predictions, we identified beliefs about toxins in vaccines as a potentially useful site of intervention–it is among the most strongly related and it is a concrete belief that could be addressed directly.Using information from the University of Oxford's "Vaccine Knowledge Project," the U.S. Department of Health and Human Services, Federal Drug Administration, and CDC, we developed an educational intervention aimed at alleviating these concerns and tested its efficacy in two experiments.

Participants.

We recruited 1491 people to participate in a two-part study via mTurk in April (Experiment 3A) and June (Experiment 3B) of 2019. One day after their initial participation, participants who completed Phase 1 and passed attention checks were invited via email to participate in Phase 2 of the study, with 1077 participants returning (1379 invited, retaining 78.1% of the Phase 1 sample). Participants were paid $1.35 for their participation in each phase of the study and were allowed to participate in Phase 2 up to two days after their initial participation in Phase 1. 51 participants (4.7%) failed at least one attention check in Phase 2 and were excluded from further analysis. This left a final sample of 1026 participants.

Table 3: Bayesian regression model coefficients for model of Vaccine Intentions combining data from Experiments 3a and 3b.

Term	Estimate	$CI_{2.5\%}$	$CI_{97.5\%}$
α_1	-4.65	-6.24	-2.49
α_2	-3.52	-5.11	-1.38
α_3	-2.56	-4.15	-0.43
α_4	-1.59	-3.17	0.56
α_5	-0.63	-2.22	1.51
α_6	1.38	-0.21	3.54
β_{post}	-0.18	-0.30	-0.06
$\beta_{condition}$	0.20	-0.15	0.59
$\beta_{post:condition}$	0.32	0.15	0.49
σ_{item}	2.14	0.90	4.77
σ_{subj}	2.79	2.64	2.95

Methods.

Procedures for Studies 3a and 3b closely followed those of Study 2, but these studies tested the effects of a novel intervention aimed at dispelling concerns about toxic additives in vaccines. Full text of this intervention is available in Appendix C of these supplementary materials. In both studies, participants were randomly assigned to either a no-intervention control condition or to the novel Vaccine Ingredients intervention.

Results.

Due to the similarity of their designs, we combined data from Studies 3a and 3b for our primary analysis. However, for full transparency we first describe analyses for these studies individually. As in Study 2, we conducted a mixed-effects Bayesian ordinal regression to assess the effects of the Vaccine Ingredients intervention on participants' *Vaccination Intentions* scores. The Vaccine Ingredients intervention improved participant's vaccination intention scores relative to a no-intervention control condition. Tables below show the results of these regression analyses for Studies 3a, 3b, and the two studies combined. In each analysis, the coefficient representing the interaction between phase and condition is positive and credibly differs from zero, indicating that the Vaccine Ingredients intervention positively impacted vaccination intentions.

However, it is worth noting that reliable changes were also observed from pretest to posttest in the control conditions in both experiments (indicated by a credibly non-zero phase coefficient in each corresponding table). Observing this in Study 3a, we suspected that these attitudes were also being influenced by other factors during the course of the study. Approximately two months later we conducted Study 3b. However, we found similar patterns of changes in this later study as well. We now speculate that news of an ongoing measles outbreak at the time of these studies (Spring of 2019) may have influenced participants' beliefs about vaccines, although it is not obvious what specific news stories could have caused these shifts from one day to the next.

We also analyzed each experiment individually.

To model the effect of the Vaccine Ingredients intervention, we followed the same procedure as in Study 2, but this time augmented the model with a node representing the Vaccine Ingredients intervention as a child of the *vaccine toxicity* belief node. To simulate the impacts of this evidence, we supplied virtual evidence to the *vaccine toxicity* node by adding a child node to the cognitive-model network with a CPT capturing this evidence ratio.

After our initial preregistration for Study 3a and 3b, we did slightly modify our model-fitting approach for the data in Study 1. This means that the final model being compared differs slightly from the exact

Table 4: Bayesian regression model coefficients for model of Vaccine Intentions in Experiment 3a.

Term	Estimate	$CI_{2.5\%}$	$CI_{97.5\%}$
α_1	-4.85	-6.46	-2.61
α_2	-3.58	-5.17	-1.32
α_3	-2.57	-4.14	-0.31
α_4	-1.61	-3.18	0.65
α_5	-0.54	-2.13	1.71
α_6	1.55	-0.04	3.80
β_{post}	-0.21	-0.38	-0.04
$\beta_{condition}$	0.37	-0.19	0.92
$\beta_{post:condition}$	0.31	0.08	0.55
σ_{item}	2.12	0.89	4.81
σ_{subj}	2.91	2.68	3.15

Table 5: Bayesian regression model coefficients for model of Vaccine Intentions in Experiment 3b.

Term	Estimate	$CI_{2.5\%}$	$CI_{97.5\%}$
α_1	-4.47	-6.03	-2.34
α_2	-3.47	-5.03	-1.36
α_3	-2.56	-4.13	-0.43
α_4	-1.57	-3.15	0.56
α_5	-0.72	-2.29	1.41
α_6	1.23	-0.33	3.37
β_{post}	-0.15	-0.32	0.02
$\beta_{condition}$	0.05	-0.45	0.54
$\beta_{post:condition}$	0.33	0.09	0.57
σ_{item}	2.14	0.90	4.73
σ_{subj}	2.69	2.47	2.91

numerical predictions of the preregistration.

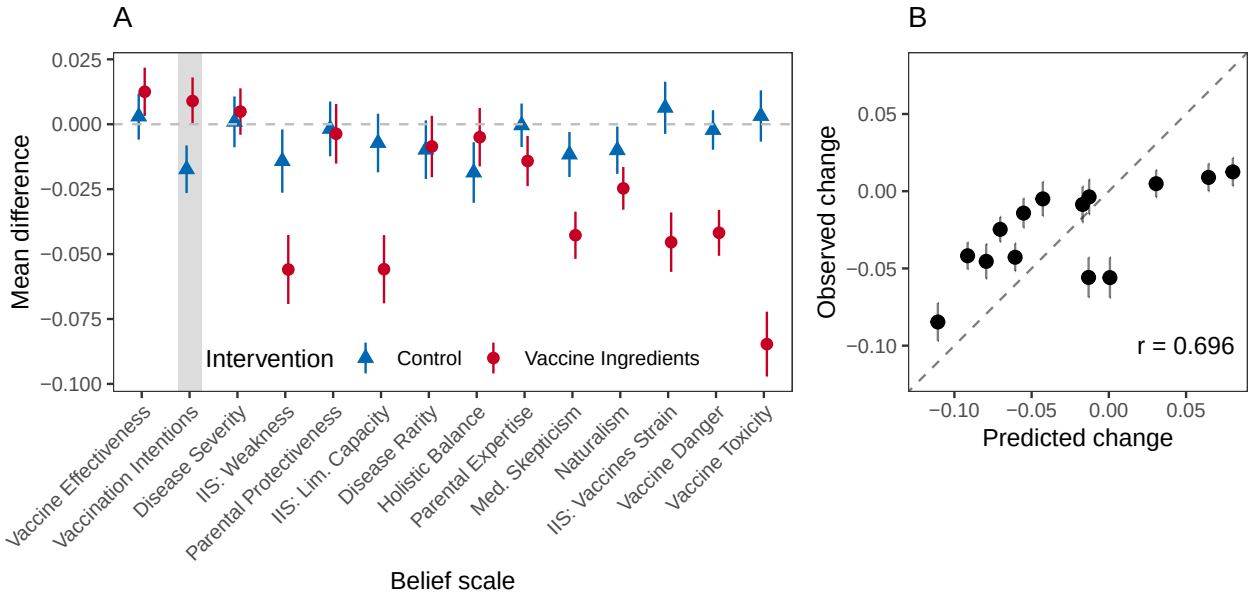

Figure 9: Belief changes following the intervention in Experiments 3a and 3b. Left, mean difference scores from pretest to posttest for the 14 belief scales across experimental conditions. Belief scales are ordered by model-predicted changes based on the observed effect on disease severity. Our primary outcome of interest, vaccination intentions, is highlighted in gray. Right, mean observed and model predicted change scores for the Vaccine Ingredients intervention condition. In both plots, error bars represent 95% confidence intervals.

Figure 9 (right) shows the correlation between predicted and observed belief changes conditional on the Vaccine Ingredients intervention. The model's predictions are significantly related to the observed changes, but the correlations are relatively weaker ($r = 0.696$) than those observed in our other studies.

In Study 3a, we observed smaller differences than predicted for vaccine effectiveness (observed minus predicted difference: -0.07), IIS: weak (-0.06), and vaccine intentions (-0.06); and larger differences than predicted for vaccine danger (+0.06), and parental expertise (+0.05); all other discrepancies were smaller than ± 0.05.

In Study 3b, we again observed smaller differences than predicted for vaccine effectiveness (observed minus predicted difference: -0.06) and vaccine intentions (-0.05); and in this case we observed larger differences than predicted for naturalism (+0.05); all other discrepancies were smaller than $\pm$ 0.05.

Study 4

Participants.

We attempted to contact and re-recruit as many of the 1129 participants who completed Study 1 to participate in a follow-up study in May of 2019. We were able to re-recruit 549 participants and, of these, 530 successfully completed the study including passing all attention checks.

Results.

Mirroring our analyses of interventions, we regressed responses to the *vaccine intentions* scale items on predictors for phase (pre-outbreak in August and September 2017 vs. post-outbreak May 2019, dummy-coded), news exposure (dummy-coded with non-exposure as the baseline), and an interaction between phase

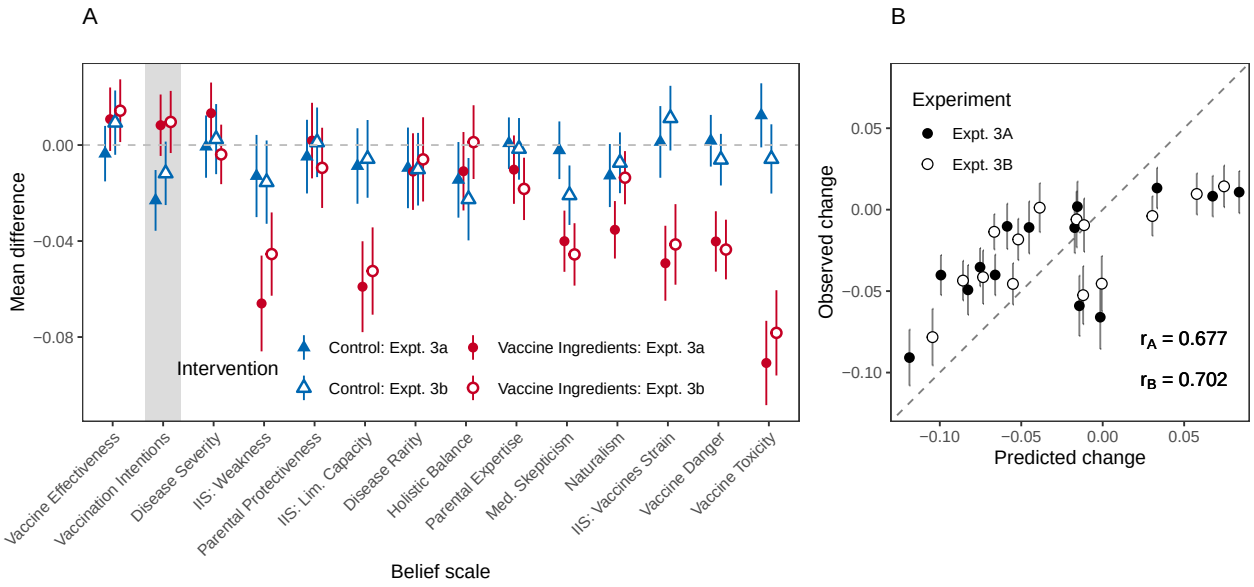

Figure 10: Belief changes following the intervention in Experiments 3a and 3b, broken out by experiment. Left, mean difference scores from pretest to posttest for the 14 belief scales across experimental conditions; our primary outcome of interest, vaccination intentions, is highlighted in grey Right, mean observed and model predicted change scores for the Vaccine Ingredients intervention condition. In both plots, error bars represent 95% confidence intervals.

Table 6: Bayesian regression model coefficients for model of Vaccine Intentions in Study 4.

Term	Estimate	$CI_{2.5\%}$	$CI_{97.5\%}$
α_1	-3.41	-4.86	-1.33
α_2	-2.48	-3.90	-0.40
α_3	-1.65	-3.08	0.43
α_4	-0.59	-2.01	1.49
α_5	0.39	-1.04	2.48
α_6	2.15	0.73	4.24
β_{phase}	0.27	0.05	0.48
$\beta_{awareness}$	1.02	0.52	1.54
$\beta_{phase:awareness}$	0.48	0.22	0.74
σ_{item}	1.87	0.76	4.33
σ_{subj}	2.44	2.26	2.64

and news exposure onto participants' scores, using a Bayesian mixed-effects ordinal regression including random intercepts by subject and item. To assess the effects of recent news exposure on *vaccine intentions*, we examined the interaction between phase and news exposure. Participants' vaccine intentions were increased by recent news exposure compared to their responses 18 months ago (see Table 6).

To model the effects of news exposure we followed the same procedure as in previous studies, but here augmented the model with a node representing news exposure as a child of the *vaccine intentions* node. To simulate the impacts of this evidence, we supplied virtual evidence to the *vaccine intentions* node by adding a child node to the cognitive-model network with a CPT capturing this evidence ratio.

Figure 11 (right) shows the correlation between predicted and observed belief changes for participants who were aware of the New York measles outbreak in May 2019.

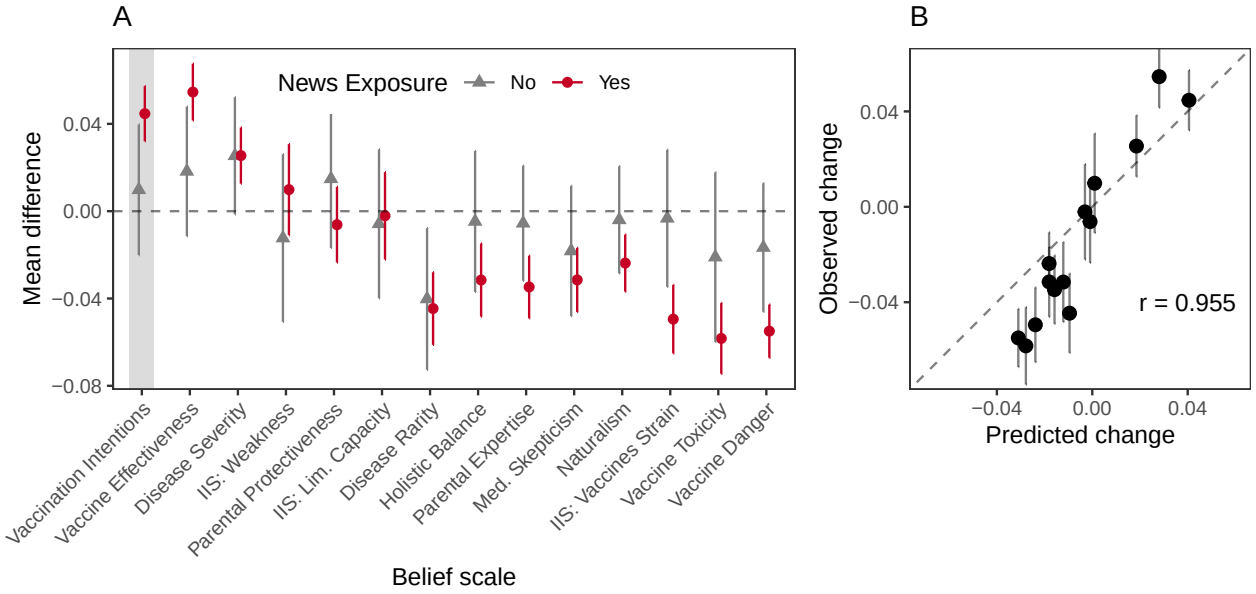

Figure 11: Belief changes following the intervention in Study 4. Left, mean difference scores from pretest to posttest for the 14 belief scales among participants who did and did not report exposure to news of the 2019 Measles outbreak in New York; our primary outcome of interest, vaccination intentions, is highlighted in gray. Right, mean observed and model predicted change scores for those who were exposed to news of the outbreak. In both plots, error bars represent 95% confidence intervals.

Participant demographic summary

The table below presents demographic information for participants across all five studies.

	Study 1	Study 2	Study 3A	Study 3B	Study 4	Overall
	(N=1129)	(N=493)	(N=515)	(N=510)	(N=549)	(N=3196)
Gender						
Female	556 (49.2%)	252 (51.1%)	232 (45.0%)	253 (49.6%)	261 (47.5%)	1554 (48.6%)
Male	573 (50.8%)	239 (48.5%)	283 (55.0%)	257 (50.4%)	288 (52.5%)	1640 (51.3%)
Other or prefer not to say	0 (0%)	2 (0.4%)	0 (0%)	0 (0%)	0 (0%)	2 (0.1%)
Age						
Mean (SD)	36.1 (11.1)	36.0 (11.2)	37.0 (11.1)	36.7 (11.2)	38.6 (12.0)	36.8 (11.3)
Median [Min, Max]	33.0 [18.0, 85.0]	33.0 [18.0, 70.0]	34.0 [19.0, 75.0]	34.0 [20.0, 72.0]	35.0 [20.0, 85.0]	34.0 [18.0, 85.0]
Missing	0 (0%)	1 (0.2%)	0 (0%)	0 (0%)	0 (0%)	1 (0.0%)
Race						
White	897 (79.5%)	381 (77.3%)	391 (75.9%)	382 (74.9%)	441 (80.3%)	2492 (78.0%)
Black or African American	75 (6.6%)	36 (7.3%)	49 (9.5%)	50 (9.8%)	30 (5.5%)	240 (7.5%)
Hispanic/Latino	57 (5.0%)	32 (6.5%)	33 (6.4%)	31 (6.1%)	25 (4.6%)	178 (5.6%)
Hawaiian or PI	0 (0%)	1 (0.2%)	2 (0.4%)	2 (0.4%)	0 (0%)	5 (0.2%)
Asian	79 (7.0%)	34 (6.9%)	31 (6.0%)	40 (7.8%)	42 (7.7%)	226 (7.1%)
Native American	5 (0.4%)	2 (0.4%)	2 (0.4%)	0 (0%)	3 (0.5%)	12 (0.4%)
Other or prefer not to say	0 (0%)	6 (1.2%)	7 (1.4%)	5 (1.0%)	0 (0%)	18 (0.6%)
Missing	16 (1.4%)	1 (0.2%)	0 (0%)	0 (0%)	8 (1.5%)	25 (0.8%)
Religion						
Buddhist	13 (1.2%)	0 (0%)	7 (1.4%)	7 (1.4%)	7 (1.3%)	34 (1.1%)
Christian	550 (48.7%)	0 (0%)	251 (48.7%)	244 (47.8%)	268 (48.8%)	1313 (41.1%)
Hindu	4 (0.4%)	0 (0%)	1 (0.2%)	5 (1.0%)	4 (0.7%)	14 (0.4%)
Jewish	14 (1.2%)	0 (0%)	2 (0.4%)	6 (1.2%)	12 (2.2%)	34 (1.1%)
Muslim	7 (0.6%)	0 (0%)	8 (1.6%)	5 (1.0%)	2 (0.4%)	22 (0.7%)
Non-religious	501 (44.4%)	0 (0%)	235 (45.6%)	225 (44.1%)	244 (44.4%)	1205 (37.7%)
Other or prefer not to say	40 (3.5%)	0 (0%)	11 (2.1%)	18 (3.5%)	12 (2.2%)	81 (2.5%)
Missing	0 (0%)	493 (100%)	0 (0%)	0 (0%)	0 (0%)	493 (15.4%)
Education						
No Diploma	7 (0.6%)	1 (0.2%)	1 (0.2%)	3 (0.6%)	4 (0.7%)	16 (0.5%)
High School	170 (15.1%)	61 (12.4%)	69 (13.4%)	58 (11.4%)	91 (16.6%)	449 (14.0%)
Some Undergraduate	295 (26.1%)	122 (24.7%)	97 (18.8%)	90 (17.6%)	137 (25.0%)	741 (23.2%)
Undergraduate Degree	425 (37.6%)	180 (36.5%)	190 (36.9%)	208 (40.8%)	199 (36.2%)	1202 (37.6%)
Some Graduate	75 (6.6%)	42 (8.5%)	39 (7.6%)	34 (6.7%)	30 (5.5%)	220 (6.9%)
Graduate Degree	138 (12.2%)	84 (17.0%)	108 (21.0%)	105 (20.6%)	79 (14.4%)	514 (16.1%)
Doctorate	14 (1.2%)	0 (0%)	8 (1.6%)	11 (2.2%)	8 (1.5%)	41 (1.3%)
Prefer not to say	0 (0%)	3 (0.6%)	3 (0.6%)	1 (0.2%)	0 (0%)	7 (0.2%)
Missing	5 (0.4%)	0 (0%)	0 (0%)	0 (0%)	1 (0.2%)	6 (0.2%)
Income ($)						
Less than 20,000	134 (11.9%)	65 (13.2%)	63 (12.2%)	51 (10.0%)	69 (12.6%)	382 (12.0%)
20,000 - 30,000	170 (15.1%)	61 (12.4%)	57 (11.1%)	53 (10.4%)	71 (12.9%)	412 (12.9%)
30,001 - 50,000	283 (25.1%)	112 (22.7%)	112 (21.7%)	115 (22.5%)	139 (25.3%)	761 (23.8%)
50,001 - 70,000	225 (19.9%)	107 (21.7%)	106 (20.6%)	120 (23.5%)	98 (17.9%)	656 (20.5%)
70,001 - 100,000	175 (15.5%)	73 (14.8%)	98 (19.0%)	102 (20.0%)	93 (16.9%)	541 (16.9%)
More than 100,000	121 (10.7%)	70 (14.2%)	76 (14.8%)	64 (12.5%)	64 (11.7%)	395 (12.4%)
Prefer not to say	21 (1.9%)	5 (1.0%)	3 (0.6%)	5 (1.0%)	15 (2.7%)	49 (1.5%)
Parent?						
No	604 (53.5%)	256 (51.9%)	261 (50.7%)	251 (49.2%)	279 (50.8%)	1651 (51.7%)
Yes	522 (46.2%)	237 (48.1%)	253 (49.1%)	259 (50.8%)	269 (49.0%)	1540 (48.2%)
Missing	3 (0.3%)	0 (0%)	1 (0.2%)	0 (0%)	1 (0.2%)	5 (0.2%)

Data Availability

Data and code for this project are available at: https://github.com/derekpowell/int-theory-vacc as well as at https://osf.io/sdfpj/

Preregistration information is available at: https://osf.io/uemq4

Materials A

Text of the belief scales used in Studies 1-4

At the beginning of each study, participants read the following instructions: "In this study, you will be asked to read a range of statements. Some of these statements may be controversial, and others less so. We are interested in your opinions about these issues. On each of the following screens, you'll see several statements. Please indicate how much you agree or disagree with each statement."

Participants then proceeded through 14 pages, with a different scale presented on each page. Scales, and questions within each scale, were presented in a random order for each participant. On each page, participants were asked to rate "how much you agree or disagree" on a 7-point scale from "Strongly disagree" to "Strongly agree." After completing all 14 scales, participants were asked standard demographics questions (gender, age, race/ethnicity, religious affiliation, educational attainment, annual household income) as well as a handful of questions about whether they were or were expecting to be a parent.

In Study 1, but not Studies 2-4, the study concluded with a brief statement about the safety of vaccines, as follows: "We very much appreciate hearing your opinions about these various issues. Some of these statements concerned vaccines. The overwhelming consensus in the scientific and medical communities is that vaccines are safe and effective. In case you're interested in learning more about this, below is a statement from the World Health Organization (WHO) on this topic, and a link to their website: 'Vaccination is one of the most cost-effective health interventions available, saving millions of people from illness, disability and death each year. Effective and safe vaccines, which protect against a number of serious diseases, are available and many promising new vaccines are being developed.' (WHO Website: www.who.int/immunization/diseases/en/)."

The 14 scales are presented in their entirety below.

Holistic balance, from McFadden et al.'s (2010) Complementary, Alternative, and Conventional Medicine Attitudes Scale (CACMAS) (Study 1: observed $\alpha = 0.82$)

1. The body is essentially self-healing and the task of a health care provider is to assist the healing process.
2. Physical and mental health are maintained by an underlying energy or vital force.
3. A patient's symptoms should be regarded as a manifestation of a general imbalance or dysfunction affecting the whole body.
4. Health and disease are a reflection of balance between positive life enhancing forces and negative destructive forces.

Naturalism (Study 1: observed $\alpha = 0.76$)

1. Natural things are always better than synthetic alternatives.
2. In some cases, it's fair to say that science has improved on nature. (reversed)
3. I'm always wary of any chemicals that were developed in a lab.
4. Overall, advances in chemistry and the creation of synthetic chemicals have done more harm than good for human health.
5. Childbirth should take place at home rather than in a hospital.
6. I'm not concerned about preservatives in food. (reversed)

Medical skepticism (Study 1: observed $\alpha = 0.78$)

1. Pharmaceutical companies are more interested in making money than in helping people be healthy.
2. FDA regulations ensure that approved pharmaceutical drugs are safe. (reversed)
3. Pharmaceutical companies put pressure on the FDA and CDC to suppress negative findings.
4. Most researchers are biased against the idea that medications or vaccinations could cause health problems.
5. Pharmaceutical companies don't make huge profits from the sale of vaccines. (reversed)

6. For researchers developing vaccines, safety is a top priority. (reversed)

Disease rarity (Study 1: observed $\alpha = 0.79$)

1. Diseases like measles and whooping cough are so rare that there is no real need to worry about them.
2. Nowadays parents can feel certain their children will not contract measles or mumps.
3. It's not paranoid for parents to worry about their children getting measles. (reversed)
4. It is almost impossible for a child currently living in the US to catch whooping cough.
5. Measles and whooping cough are more common than people realize. (reversed)

Disease severity (Study 1: observed $\alpha = 0.85$)

1. Measles, mumps, and whooping cough are no more severe than the flu. (reversed)
2. Childhood diseases are serious diseases.
3. Diseases like measles, mumps, and whooping cough can cause permanent damage to a child who contracts them.
4. Childhood diseases are not that big a deal. (reversed)
5. Diseases like measles and whooping cough are extremely dangerous for young children.

Vaccine effectiveness (Study 1: observed $\alpha = 0.85$)

1. Your chances of getting a disease after being vaccinated against it are incredibly low.
2. Improved nutrition and sanitation played a larger role than vaccines in the decline of diseases like measles. (reversed)
3. Recent outbreaks of diseases like measles and whooping cough show that vaccines don't work very well. (reversed)
4. Vaccines are one of the most effective medical treatments.

Vaccine danger (Study 1: observed $\alpha = 0.88$)

1. Childhood vaccines can cause autism.
2. It is safe for infants to receive multiple vaccines at once. (reversed)
3. Many children have severe adverse reactions to the MMR vaccine.
4. Vaccines are a very safe medical treatment. (reversed)

Toxic additives in vaccines (Study 1: observed $\alpha = 0.91$)

1. Vaccines do not contain dangerous amounts of toxins. (reversed)
2. Vaccines are actually highly toxic to humans.
3. Exposure to certain additives in vaccines can cause major health problems.
4. The quantities of chemical additives in vaccines are so small that they do not pose any health risks. (reversed)

Infant immune system: Vaccines strain (Study 1: observed $\alpha = 0.86$)

1. Vaccines strain the immune system much like the actual disease.
2. One great benefit of vaccines is that they provide immunity without the immune system having to work so hard. (reversed)
3. Vaccines can use up the capacity of the immune system, leaving a baby vulnerable.
4. Vaccines are not very taxing for an infant's immune system. (reversed)
5. Getting multiple vaccines at once can exhaust an infant's immune system.

Vaccination intentions (Study 1: observed $\alpha = 0.85$)

1. The risk of side effects outweighs any protective benefits of vaccines. (reversed)
2. If I had a young baby I would have him or her vaccinated.
3. If I had a baby, I would opt to spread out his or her vaccinations, rather than following the traditional schedule. (reversed)
4. I feel confident that vaccinating children is the right thing to do.
5. I would never vaccinate my child. (reversed)

Parental expertise (Study 1: observed $\alpha = 0.83$)

1. Every individual is different and a parent knows their own child best.
2. Parents have insights into their children's health and well-being that no health professional can match.
3. When a child encounters health problems, the child's parents are often better equipped than doctors to identify the many subtle causes of those issues.
4. Parents should trust a doctor's advice even if it goes against their intuitions. (reversed)
5. Advice from doctors is the best resource for keeping your child healthy. (reversed)

Infant immune system: Limited capacity (Study 1: observed $\alpha = 0.86$)

1. When babies are exposed to one illness it leaves them especially vulnerable to other viruses and infections.
2. Babies are able to fight off multiple viruses at the same time. (reversed)
3. Babies' immune systems can only develop a limited number of antibodies.
4. Babies' immune systems are severely limited in how much they can cope with at one time.
5. Babies' immune systems can deal with many threats at one time. (reversed)

Infant immune system: Weakness (Study 1: observed $\alpha = 0.86$)

1. Babies should be sheltered from exposure to viruses and bacteria to the greatest extent possible.
2. Babies have a very limited ability to fight off diseases.
3. Babies' immune systems are weak.
4. Babies have an innate ability to fight off diseases. (reversed)
5. Babies' immune systems are more robust than many people think. (reversed)

Parental protectiveness (Study 1: observed $\alpha = 0.73$)

1. Parents should strive to control their baby's environments as much as possible.
2. Parents of babies should never leave anything to chance.
3. There is no such thing as being overprotective of a baby.
4. Many things that parents of babies worry about won't actually have major consequences.
5. Some parents go overboard trying to make everything perfect for their babies. (reversed)

Appendix B

Text of the Disease Risk intervention adapted from (Horne et al., 2015). See supplemental materials for this paper to see images presented.

All children should be vaccinated for measles, mumps, and rubella. These are serious diseases. Please read the descriptions of these diseases and carefully view the pictures.

Measles

Measles virus causes rash, cough, runny nose, eye irritation, and fever. It can lead to ear infections, pneumonia, seizures (jerking and staring), brain damage, and death.

Mumps

Mumps virus causes fever, headache, and swollen glands. It can lead to deafness, meningitis (infection of the brain and spinal cord covering), painful swelling of the testicles or ovaries, and, rarely, death.

Rubella (German Measles)

Rubella virus causes rash, mild fever, and arthritis (mostly in women). If a woman gets rubella while she is pregnant, she could have a miscarriage or her baby could be born with serious birth defects.

You or your child could catch these diseases by being around someone who has them. They spread from person to person through the air.

Measles, mumps, and rubella (MMR) vaccine can prevent these diseases. Most children who get their MMR shots will not get these diseases. Many more children would get them if we stopped vaccinating.

Here is a true story that shows why vaccination is so important.

If you hear "106 degrees" you probably think "heat wave," not a baby's temperature. But for Megan Campbell's 10-month-old son, a life-threatening bout of measles caused fevers spiking to 106 degrees and sent him to the hospital. "We spent 3 days in the hospital fearing we might lose our baby boy," Campbell said. "He couldn't drink or eat, so he was on an IV, and for a while he seemed to be wasting away. When he could drink again, we got to take him home. But the doctors told us to expect the disease to continue to run its course, including high fever – which spiked as high as 106 degrees. We spent a week waking at all hours and soothing him with damp washcloths."

Thankfully, the baby recovered fully.

Megan now knows that her son was exposed to measles when another mother brought her ill son into their pediatrician's waiting room.

Materials C

Text of the Vaccine Ingredients intervention targeting beliefs about toxins in vaccines.

Vaccines are one of the most important medical advances in history, but some parents worry that vaccines contain toxins that could harm their babies. In fact, all of the ingredients in a vaccine serve important purposes that help ensure the safety and effectiveness of the vaccine. Decades of research have shown vaccine ingredients to be safe, even in much larger quantities than those found in vaccines.

What's in a vaccine?

Active ingredients

The key ingredient in all vaccines is one or more active ingredients, whose purpose is to train the immune system to recognize and combat the pathogens that cause diseases. To do this, certain molecules from the pathogen are introduced into the body to trigger an immune response. These molecules are called antigens, and they are present on all viruses and bacteria. By injecting these antigens into the body, the immune system can safely learn to recognize them as hostile invaders, produce antibodies, and remember them for the future.

Other Ingredients

A typical dose of an injected vaccines is just 0.5 millilitres of liquid, which is just a few drops. Apart from active ingredients, the main ingredient in vaccines is water. All other ingredients weigh a few thousandths of a gram or even less. All of the ingredients in vaccines have been proven to be safe for children in these minute quantities.

All of the ingredients in a vaccine serve important purposes, and are present to ensure that the vaccine is safe and effective. Some of these added ingredients are adjuvants, added to enhance the immune system response; others are antibiotics, to prevent contamination during the manufacturing process; and still other ingredients act as preservatives and stabilizers.

To take one example, consider one of the scariest things you may have heard about the ingredients in vaccines: that some vaccines contain formaldehyde. It may surprise you to learn that formaldehyde is actually produced by the human body as part of normal metabolic processes. Human bodies—even babies' bodies—naturally produce, process, and safely excrete formaldehyde. The amount of formaldehyde in vaccines is extremely small compared to what our bodies produce on their own: the amount of formaldehyde naturally present in a 2-month-old infant's blood (around 1.1 milligrams) is ten times greater than the amount found in any vaccine (less than 0.1 milligrams). In addition, many healthy foods–such as apples, grapes, bananas, and pears–naturally contain formaldehyde. A pear, for example, contains around 50 times more formaldehyde than is found in any vaccine.

Conclusion

Vaccines are extremely safe. A vaccine is just a few drops of liquid, the vast majority of which is water, and all of the ingredients in vaccines serve specific purposes to help make them safe and effective. Although some vaccine ingredients may sound scary, many of them are natural substances that are already in our bodies. In large quantities anything can be toxic, but extensive testing has shown that none of the ingredients in vaccines are toxic in the tiny quantities with which they're used. Vaccines have been carefully studied to ensure both their safety and their effectiveness in protecting babies from serious diseases.

References

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Dubé, E., Vivion, M., & MacDonald, N. E. (2015). Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: Influence, impact and implications. *Expert Review of Vaccines*, *14*(1), 99–117. https://doi.org/10.1586/14760584.2015.964212

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8

Holyoak, K. J., & Cheng, P. W. (2011). Causal Learning and Inference as a Rational Process: The New Synthesis. *Annual Review of Psychology*, *62*(1), 135–163. https://doi.org/10.1146/annurev.psych.121208.131634

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*(33), 10321–10324. https://doi.org/10.1073/pnas.1504019112

Larson, H. J., Jarrett, C., Schulz, W. S., Chaudhuri, M., Zhou, Y., Dube, E., Schuster, M., MacDonald, N. E., & Wilson, R. (2015). Measuring vaccine hesitancy: The development of a survey tool. *Vaccine*, *33*(34), 4165–4175. https://doi.org/10.1016/j.vaccine.2015.04.037

Maaskant, P. P., & Druzdzel, M. J. (2008). An independence of causal interactions model for opposing influences. *Proc. 4th European Workshop on Probabilistic Graphical Models*, 185–192.

McFadden, K. L., Hernández, T. D., & Ito, T. A. (2010). Attitudes Toward Complementary and Alternative Medicine Influence Its Use. *EXPLORE*, *6*(6), 380–388. https://doi.org/10.1016/j.explore.2010.08.004

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455–485. https://doi.org/10.1037/0033-295X.111.2.455

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann.

Quine, W. V. (1951). Two Dogmas of Empiricism. *The Philosophical Review*, *60*(1), 20–43. https://doi.org/10.2307/2181906

Quine, W. V., & Ullian, J. S. (1978). *The web of belief* (Second edition). Random House.

Salmon, D. A., Dudley, M. Z., Glanz, J. M., & Omer, S. B. (2015). Vaccine Hesitancy. *American Journal of Preventive Medicine*, *49*(6), S391–S398. https://doi.org/10.1016/j.amepre.2015.06.009

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *arXiv:0908.3817 [Stat]*. http://arxiv.org/abs/0908.3817

Weisman, K., & Markman, E. M. (2017). Theory-based explanation as intervention. *Psychonomic Bulletin & Review*, *24*(5), 1555–1562. https://doi.org/10.3758/s13423-016-1207-2

Williams, S. E. (2014). What are the factors that contribute to parental vaccine-hesitancy and what can we do about it? *Human Vaccines & Immunotherapeutics*, *10*(9), 2584–2596. https://doi.org/10.4161/hv.28596

Williamson, J. (2001). Bayesian Networks for Logical Reasoning. In C. Gomez & T. Walsh (Eds.), *Proceedings of the AAAI Fall Symposium on using Uncertainty within Computation* (pp. 136–143).