Articulating lay theories through graphical models: A study of beliefs surrounding vaccination decisions

Derek Powell
derekpowell@stanford.edu
Department of Psychology
Stanford University

Kara Weisman
kweisman@stanford.edu
Department of Psychology
Stanford University

Ellen M. Markman
markman@stanford.edu
Department of Psychology
Stanford University

Abstract

How can we leverage the cognitive science of lay theories to inform interventions aimed at correcting misconceptions and changing behaviors? Focusing on the problem of vaccine skepticism, we identified a set of 14 beliefs we hypothesized would be relevant to vaccination decisions. We developed reliable scales to measure these beliefs across a large sample of participants ($n = 1130$) and employed state-of-the-art graphical structure learning algorithms to uncover the relationships among these beliefs. This resulted in a graphical model describing the system of beliefs relevant to childhood vaccinations, with beliefs represented as nodes and their interconnections as directed edges. This model sheds light on how these beliefs relate to one another and can be used to predict how interventions aimed at specific beliefs will play out across the larger system. Moving forward, we hope this model will help guide the development of effective, theory-based interventions promoting childhood vaccination.

Keywords: graphical modeling; lay theories; conceptual change; behavioral interventions

Much of the richness of human thought depends on our ability to combine and synthesize information into coherent belief systems, lay theories, and mental models. These cognitive processes are vital for interpreting, explaining, and predicting events; and for planning actions to intervene on the course of these events. But these same abilities can sometimes lead people astray, generating misconceptions that result in inappropriate and even dangerous actions. Here, we focus on one striking and timely example: The resurgence of diseases like measles in the wake of widespread misconceptions about the safety of childhood vaccines.

In a larger project, we aim to develop effective ways to address this and other misconceptions by leveraging the cognitive science of lay theories to effect conceptual and behavioral change (see Weisman & Markman, 2017, for a review of this approach). In this paper, our goal is to enrich our understanding of the conceptual "ecosystem" that supports or discourages vaccination. To this end, we develop a graphical model that describes the system of beliefs relevant to childhood vaccinations, representing these beliefs as nodes and their interconnections as directed edges. Moving forward, we hope the use of these formal techniques will let us make quantitative inferences and predictions to help guide the development of educational interventions.

Vaccine beliefs and misconceptions

In the early 2000s, now-discredited research led many people to believe that childhood vaccinations, such as the Measles, Mumps, and Rubella (MMR) vaccine, could increase children's risk for autism. Vaccination rates declined in many communities, leading to a resurgence of preventable childhood diseases: In 2014 the CDC tracked 667 cases of measles in the US, where the disease had previously been eradicated (CDC, 2015). Vaccines do not, in fact, cause autism (Taylor, Swerdfeger, & Eslick, 2014), but these misconceptions have proved to be remarkably difficult to correct (e.g., Betsch & Sachse, 2013; Horne, Powell, Hummel, & Holyoak, 2015).

One challenge to addressing misconceptions is that they are often embedded in larger, internally coherent belief systems that guide how people interpret and respond to evidence (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). Suppose one thinks the infant immune system is immature, weak, and easily overwhelmed: It might then seem unreasonable to vaccinate a 2-month-old baby against 5 or more diseases at once, as the CDC recommends. Similarly, if someone believes that the medical community is unduly influenced by pharmaceutical companies, she might be skeptical when medical studies come out in favor of these companies' interests. Such beliefs might sustain the misconception that vaccinating children is dangerous, even in the face of counter-evidence.

For educational interventions to be effective, they must be sensitive to the broader conceptual context in which they'll be interpreted. In the case of vaccine attitudes, interventions that simply emphasize the safety of vaccines may not be convincing to people who hold strong beliefs about the vulnerability of the infant immune system or corruption in medical research—but other beliefs might be more amenable to revision. Consistent with this, Horne et al. (2015) found that straightforward reassurances of vaccine safety were ineffective in changing people's attitudes toward vaccination, but informing people about the risks of measles, mumps, and rubella resulted in more positive views of childhood vaccination. As in many domains, our understanding of the conceptual system driving vaccination decisions is limited. Having hypothesized that some set of beliefs might be relevant to people's vaccination decisions, it would be extremely useful to validate these intuitions and specify precisely how these beliefs relate to or inform one another.

How can we effectively transform a qualitative account into a useful, testable, model of a lay theory? In this paper, we describe a graphical modeling approach to developing a rich, formal theory of the beliefs surrounding vaccination decisions. We began by identifying potentially relevant beliefs, developing reliable instruments for measuring them, and using those instruments to survey a large sample of participants. Then, we used a state-of-the-art graphical modeling

approach—Bayesian network structure learning (for a review, see Scutari & Denis, 2014)—to discover and describe connections among these beliefs and represent them in a quantitative model. We consider this project a first step in a longer process that, we hope, will yield a rich, quantitatively precise theory of this conceptual system.

Study

Our goal was to use behavioral data to develop a graphical model of a conceptual system that could support or discourage vaccination. This process involves many choices about data representation, as well as trade-offs between the fit, complexity, and intelligibility of the models produced. Here we describe the steps we took to build this model, highlighting key decision points that shaped the resulting model.

Scale development: Identifying and measuring relevant beliefs

The first and perhaps most consequential decision points concern the set of beliefs we hypothesized are relevant to vaccination decisions, and how we chose to measure these beliefs. Our outcome of interest was participants' intentions to vaccinate their children (*vaccination intentions*). Drawing on a variety of sources, including research on anti-vaccine skepticism, anti-vaccine websites, and our own qualitative surveys with vaccine skeptics (not reported here), we generated a list of 13 underlying beliefs that might influence this outcome.

These included two broad worldviews: (1) *naturalism*, a general preference for natural over artificial things; and (2) *holistic balance*, one important aspect of attitudes toward alternative medicine (McFadden, Hernndez, & Ito, 2010), as well as three slightly more specific theories about parenting and medicine: (3) general *parental protectiveness*; (4) *parental expertise*, namely the belief that parents usually know more about their children's health than medical experts; and (5) *medical skepticism*, including concerns about pharmaceutical companies and corruption in the medical community. In addition, we identified a variety of specific claims about vaccines that seemed important to people's arguments for and against vaccination, including beliefs about (6) the overall safety of vaccines (*vaccine danger*); (7) *toxic additives in vaccines*; and (8) *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 (9) *disease rarity* and (10) *disease severity*. Beyond this, we theorized that intuitive theories of the infant immune system might be relevant, including beliefs that (11) the infant immune system is weak (*IIS: weakness*); (12) the infant immune system is limited in its capacity and can be easily overwhelmed (*IIS: limited capacity*); and (13) vaccines strain the infant immune system (*IIS: vaccines strain*).

We then developed psychometrically robust scales to measure these beliefs, stipulating 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 (Cronbach's $\alpha \geq .80$). After extensive piloting and refinement, we created 14 scales that met these criteria, including one preexisting scale (the "holistic balance" subscale from McFadden et al., 2010). Final observed reliability ranged from .73 to .91. (A full list of items for all scales is available at `https://osf.io/dc5j8/`.)

Method

To investigate relationships among the beliefs surrounding vaccination intentions, we examined covariation among these beliefs across a large sample of participants. For instance, if someone strongly endorses medical skepticism, how might this influence their beliefs about the toxicity of vaccines, or the severity of diseases like measles? These observed covariation relationships shed light on how these beliefs hang together and influence one another and, combined with structure learning algorithms, provide a path toward approximating this conceptual system.

Participants 1200 people participated via Amazon Mechanical Turk. 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. Participants were paid $1.60 for about 8 minutes of their time. Repeat participation was prevented.

Procedure Participants were told that we were interested in their opinions about a variety of topics. They then proceeded through our 14 scales, rating each statement on a scale from "Strongly disagree" (coded as -3) to "Strongly agree" (+3); the order of presentation of these scales and the order of questions within each scale was randomized for each participant.

Two attention checks (e.g., "Please select somewhat agree") were embedded randomly among these questions; the 70 participants who failed at least one of these checks were excluded from further analyses. This left a final sample of n = 1130 (94% of our full sample).

Data preparation Scores for each scale were calculated as the average of the responses to questions in that scale, after reverse-coding; for all scales, the theoretical range of scores was -3 to +3. The final dataset for modeling included 14 scores for each participant.

Model Building

Our primary goal was to build a formal model that could approximate the conceptual relationships among beliefs related to vaccine intentions. This bring us to our second decision point: How to model the data. We conceived of these beliefs as influencing one another, where influences are directed from one belief to another as in, for example, causal relationships (Pearl, 2000) and logical implication (Williamson, 2001). These types of asymmetric relations can be well-captured in a Directed Acyclic Graph (DAG), where each belief is represented as a node in a network, and all edges between nodes are directed, i.e., connections run in one direction only. For instance, an edge from naturalism to medical

skepticism would indicate that naturalism beliefs influence medical skepticism. Because we measured beliefs continuously, we employed gaussian (linear) DAGs.

This class of models has several desirable qualities. First, there are efficient algorithms for "learning" these network structures from data, allowing us to discover possible relationships among beliefs using observed correlations in a large sample of participants. Second, a DAG can be used to generate inferences based on information about a subset of the network's nodes. This allows us to predict a person's beliefs about a given topic (e.g., vaccine safety) based on observations of their beliefs about another topic (e.g., medical skepticism). Finally, these networks are capable of generating predictions about the consequences of intervening on nodes within these systems, an important advantage when using these networks to craft real-world interventions.

Incorporating theory

Structure-learning algorithms operate in a "bottom-up" fashion, generating a model based on raw data. Still, there are opportunities to exert "top-down" influences on this theory-building process. This brings us to our third decision point: whether and how to constrain the search for the structure connecting these beliefs. By "whitelisting" or "blacklisting" connections between nodes, we can stipulate that they must or must not be included in the final model. Such constraints could be specific (e.g., a link from A to B must be included) or broad (e.g., C has no "parents," i.e., no incoming connections; D has no "children," i.e., does not feed into any other nodes).

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 measured, and labeled these "worldviews"; we considered our outcome of interest, *vaccine intentions*, to be the most concrete measurement of a specific "intention." Figure 3 shows the level assigned to each node in the network. We used this hierarchy of beliefs to induce a blacklist that would constrain our search space. We made the assumption that the beliefs surrounding vaccine decisions would be best described as a generative model, in which more abstract beliefs set expectations for more concrete beliefs or observations (following, e.g., Jern, Chang, & Kemp, 2014). In other words, "worldviews" could feed directly into "theories," "claims," or "intentions," but none of these more concrete beliefs could feed into "worldviews"; likewise, "theories" could feed into "claims" or "intentions" (but not vice versa); and "claims" could feed into "intentions" (but not vice versa). This approach offers a highly generalizable means to incorporating existing a priori theories into the structure learning process.

Structure learning algorithms

We now turn to our fourth decision-point, the selection of a structure-learning algorithm. Here, we consider two structure learning algorithms implemented in the bnlearn R package (v4.2)– the score-based hill climbing (HC) algorithm and the hybrid Min-Max Hill Climbing (MMHC) algorithm (Scutari, 2010). In addition, we introduce our own hybrid approach that may offer some appealing qualities for our purposes. Our approach is similar to MMHC, which first restricts the search space for a directed graph by finding an undirected "skeleton" describing conditional-independence relationships among variables. However, unlike the MMHC algorithm, which uses the "min-max parents" (MM) heuristic algorithm to constrain the search space, we use state-of-the-art Bayesian structure learning algorithms implemented in the BDgraph R package (Mohammadi & Wit, 2017) to identify this undirected skeleton. Like MMHC, our approach then uses the HC algorithm to find a directed graph. We will refer to this custom algorithm as "BDHC."

Achieving intelligibility

Because we aim to develop interventions based on the theory emerging from our model, an important desiderata for this model is intelligibility. This raises a fifth decision point: the degree to which we are willing to trade off predictive accuracy in exchange for greater intelligibility.

Some degree of simplicity is likely necessary for intelligibility. One proxy for simplicity is sparsity, or the number of edges present in the graph. Both MMHC and our custom algorithm, BDHC, offer a fairly direct means to impose varying degrees of sparsity on the resulting graph. In MMHC the modeler is free to choose the (frequentist) α criterion for the restriction phase: A higher α value results in fewer edges. Similarly, using BDHC the modeler can set the threshold for the posterior probability of edges to be included in the skeleton: In this approach, edges are present in the final graph only when the posterior probability that there is a dependency between these nodes, independent of the other variables, is greater than some specified threshold (e.g., .95).

Cross-validation and algorithm selection

We have highlighted five key decision points in constructing our model. Several of these, including choosing an algorithm and a threshold for retaining edges, can be aided by empirical cross-validation procedures, which allow us to explore a large space of models while avoiding overfitting. With this in mind, we split our data into a "training split" (80% of the data), which we used to develop and compare models, and a "testing split" (20%), which we used to validate the final model's performance. We performed 10 runs of 10-fold cross-validation on the training data to compare the performance of our different approaches, using identical fold-splits for all models. Using this procedure, we compared the HC, MMHC, and BDHC algorithms, using various values for alpha (MMHC) and posterior thresholds (BDHC), and including or omitting our theory-based blacklist.

Cross-validation results comparing these models are shown in Figure 1. We were interested both in how well the models produced by these algorithms performed in an out-of-sample prediction (as indexed by their log likelihood loss) as well as

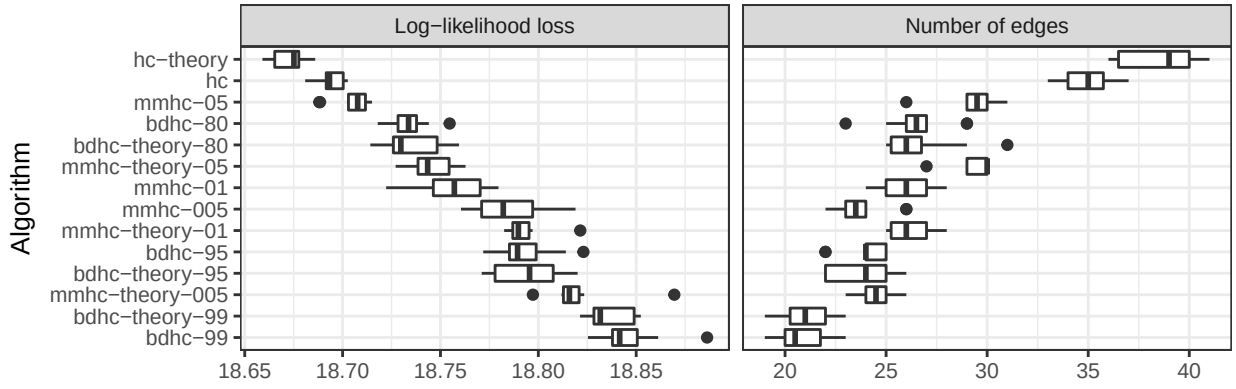

Figure 1: Cross-validation results. Left: Log-likelihood loss predicting out-of-sample data across 10 run 10-fold cross-validation. Right: Number of edges in models generated by each algorithm. Algorithms are named according to the use of the theory-based blacklist, and the threshold used (e.g., mmhc-theory-05 is the MMHC algorithm with the theory-based blacklist and $\alpha = .05$).

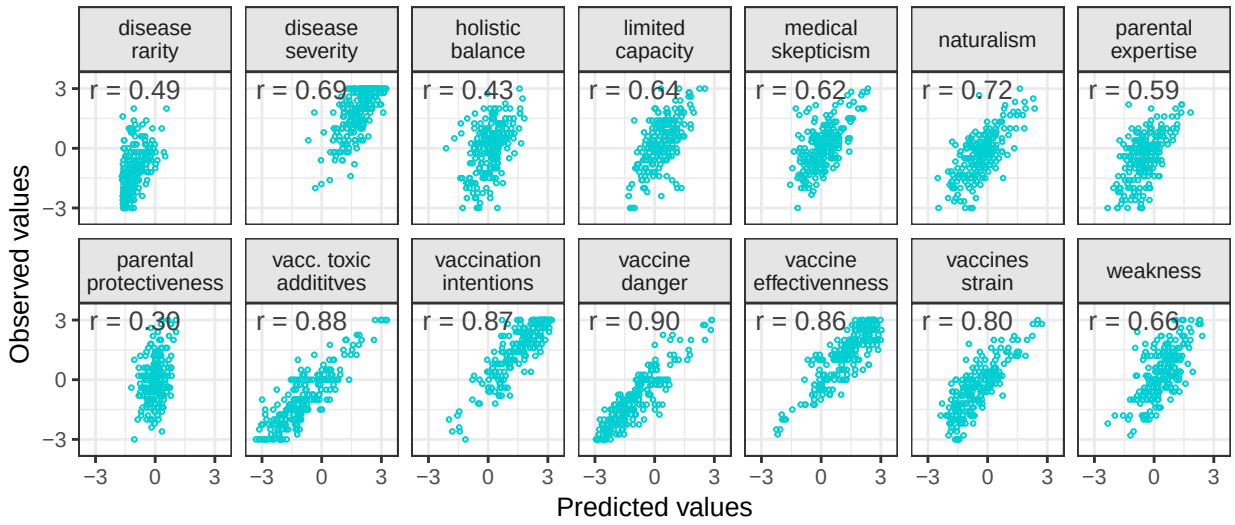

Figure 2: Observed versus predicted values for each belief in the testing set, with predictions from the final BDHC model using posterior probability threshold = .95 and fit to the training split.

in how complex the resulting models were (as indexed by the number of edges in the resulting graphs).

A few points are apparent from the results of cross-validation. First, the inclusion of the theory-based blacklist ("tiers" of abstractness) had relatively little impact on model performance. This is promising, as it suggests our existing theory is not in conflict with the data.

Second, there is a trade-off between the degree to which the algorithm is tuned toward sparsity and the resulting fit, such that more complex models generally provide somewhat better fits. If we were prioritizing predictive power, we would proceed with the best-fitting model (HC-theory); if we were prioritizing simplicity, we might opt to proceed with the sparsest model (BDHC-theory-99). For the purposes of designing real-world interventions, we would like a model that both allows us to make accurate out-of-sample predictions and that provides an intelligible theory. Striking the "right" balance between predictive power and intelligibility is difficult to resolve formally.

We thus proceeded informally, attempting to balance concerns for fit and intelligibility in proportion to our project's goals. Averaging across folds, the likelihood ratio of observed data under the best-fitting model (HC-theory) compared with the worst-fitting model (BDHC-theory-99) was only 1.19. Although reliable, these differences in fit are not sufficient to motivate adopting the most complex models. Instead, we sought to identify the best-fitting model that was sufficiently simple and intelligible for our purposes. To assess intelligibility more directly, we used each algorithm to learn a graph based on the entire set of training data (n = 904). From among these different options, we chose to proceed with the model resulting from the BDHC method with a posterior probability threshold of .95.

This resulted in a partially-directed acyclic graph (PDAG)

with three undirected edges. To generate model predictions for validation, we chose to set these edge directions arbitrarily, under the assumption that they will not meaningfully impact prediction performance due to score-equivalence (Scutari & Denis, 2014). The final resulting network is shown in Figure 3.

Validating the model's performance

To evaluate the model's performance, we tested its accuracy in predicting responses among the remaining 20% testing split (n = 226). After learning the network and fitting its parameters using the training data split, we generated predictions for held-out participants' responses for each variable by conditioning the network on the remaining 13 (observed) variables. Figure 2 compares the model's predictions with participants' actual responses.

Collapsing across all variables, the average correlation between predicted and observed responses was r = .825, accounting for 68.1% of the variance in observed responses. Correlations between observed and predicted values ranged from .304 to .899 across the different belief scales. In general, the model shows greater predictive accuracy for more central beliefs (e.g., *vaccine danger*) than for more distant beliefs (e.g., *parental protectiveness*). Altogether, this out-of-sample predictive performance suggests this model can usefully predict and explain participants' beliefs.

Discussion

We developed a graphical model of a conceptual "ecosystem" surrounding vaccination decisions, by combining an initial qualitative theory with behavioral data using Bayesian network structure learning. The resulting model (Figure 3) offers a preliminary description of the conceptual systems that support and discourage vaccination decisions.

The ultimate value of this model rests heavily on its validation by future interventional studies. With that in mind, we consider some preliminary insights and implications. First, this model confirms that the 14 beliefs we hypothesized would be relevant to vaccine decisions are, in fact, closely related to each other and to participants' intentions to vaccinate their children. Many of the conceptual connections revealed by this model make sense intuitively. For example, beliefs about the effectiveness of vaccines, the safety or danger of vaccines, and the severity of childhood diseases are the three nodes with direct connections to *vaccination intentions*. Such findings provide one check on the success of the model-building process, and suggest it is uncovering meaningful relationships.

Other findings may shed new light on the role of lay theories in vaccine decisions. For example, a "naturalist" worldview—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 finding supports some of our earlier speculations as to why interventions have often failed to alter vaccine skepticism: These beliefs may be tied into far-ranging worldviews that affect many aspects of people's thinking, including their interpretation and response to evidence about the safety of vaccines.

Finally, the current model highlights certain beliefs that might be especially influential in shaping vaccination decisions, such as beliefs about *naturalism*, *vaccine danger*, *vaccine effectiveness*, and *toxic additives in vaccines*. Of course, some of these beliefs may be more or less amenable to interventions. For instance, previous work suggests that it may be difficult to craft interventions that effectively dispel beliefs about *vaccine danger* (e.g., Horne et al., 2015). Still, by revealing the interconnections among these beliefs, the model suggests ways to overcome these challenges. One promising approach could be to combine successful interventions from past research, such as providing information about the severe dangers of diseases like measles for infants and young children (Horne et al., 2015), with information about how and why vaccines work so well to protect children from these diseases (targeting *vaccine effectiveness*).

Conversely, some interventions that initially seemed promising now seem more complicated. For example, we initially hoped that providing information to parents about how the infant immune system works—in particular, dispelling the misconception that it has a limited capacity—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 = -.097 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, which might, in turn, *encourage* vaccination. In light of this, we speculate that attempting to dispel beliefs about limited capacity might have no effect on a person's vaccine intentions (due to these countervailing forces)—or such an intervention might have different effects for different people, depending on their auxiliary beliefs (e.g., about disease severity). Simulation studies using this model could help elucidate these possibilities, and will be critical as we continue to pursue effective interventions.

Moving forward, we envision an iterative process in which we continue to combine bottom-up, data-driven insights with top-down theorizing to refine our understanding and develop effective interventions. First, we can use the model to simulate how interventions targeting specific beliefs or combinations of beliefs will affect beliefs throughout the wider network. Based on these predictions, we can choose optimal sites of intervention, craft interventions aimed at changing these target beliefs, and measure their effects. Studies and simulations will allow us to identify where the model succeeds or fails, and revise our model and theory accordingly (e.g., by reversing the direction of edges, adding missing vari-

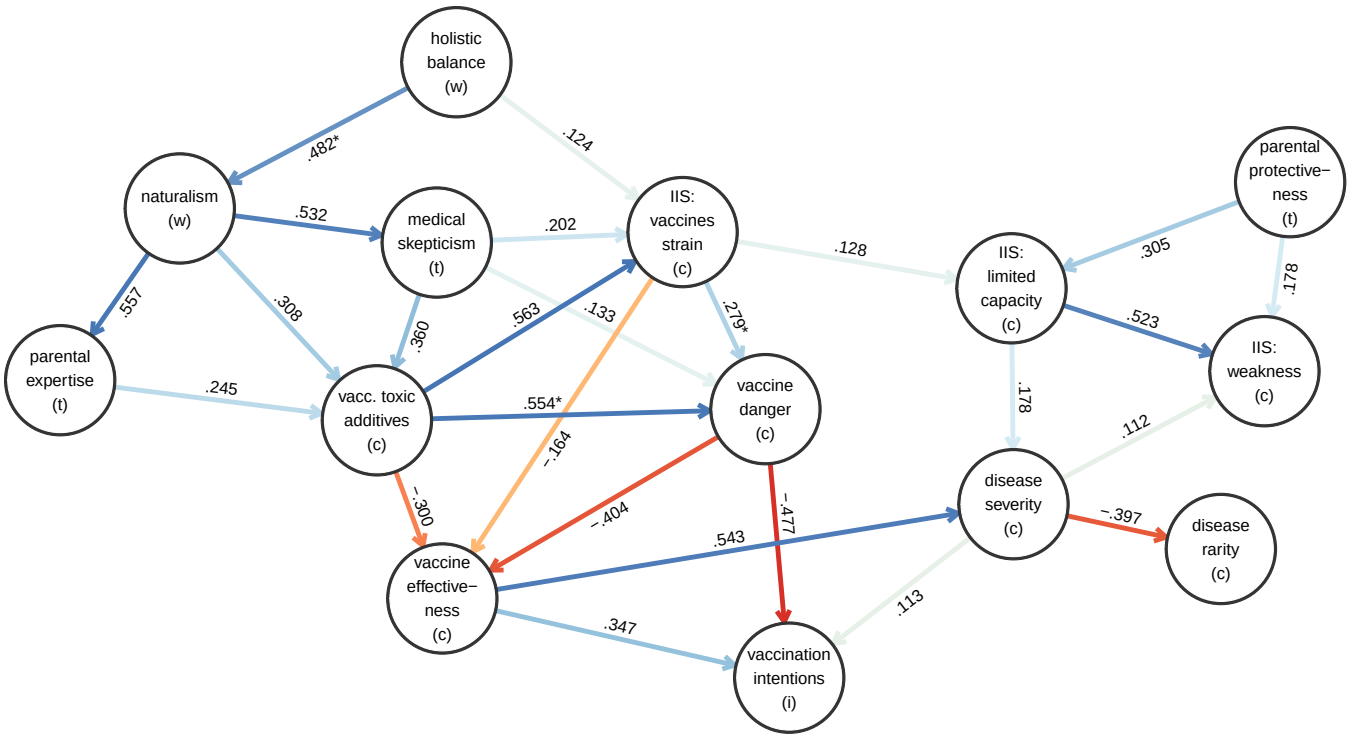

Figure 3: Final BDHC model using posterior probability threshold = .95. Nodes are labeled for abstractness, from worldviews (w), to theories (t), claims (c), and intentions (i). Edge weights indicate standardized linear coefficients from the gaussian model, which can be interpreted as regression coefficients. Asterisks indicate edges that were directed arbitrarily.

ables, specifying interactions, or modeling non-linear relationships). If these interventions have positive outcomes, we can begin translating them into more applied contexts.

Developing educational interventions is difficult, and testing these interventions, particularly in person, can be extremely costly. Here, we illustrated a promising and novel method for moving effectively from intuitions about lay theories to empirically validated methods for correcting misconceptions and improving decisions.

References

Betsch, C., & Sachse, K. (2013). Debunking vaccination myths: strong risk negations can increase perceived vaccination risks. *Health Psychology*, *32*(2), 146–55.

Centers for Disease Control and Prevention. (2015). Measles in the us. Retrieved from `http://www.cdc.gov/media/releases/2015/t0129-measles-in-us.html`

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(33), 10321–4.

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

McFadden, K. L., Hernndez, T. D., & Ito, T. A. (2010). Attitudes toward complementary and alternative medicine influence its use. *EXPLORE: The Journal of Science and Healing*, *6*(6), 380–388.

Mohammadi, A., & Wit, E. C. (2017, August). BDgraph: Bayesian structure learning in graphical models using birth-death mcmc. R package version 2.40.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York, NY, USA: Cambridge University Press.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package, *35*(3).

Scutari, M., & Denis, J.-B. (2014). *Bayesian networks with examples in R*. Boca Raton: Chapman & Hall.

Taylor, L. E., Swerdfeger, A. L., & Eslick, G. D. (2014). Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies. *Vaccine*, *32*(29), 3623–3269.

Weisman, K., & Markman, E. M. (2017). Theory-based explanation as intervention. *Psychonomic Bulletin and Review*, *24*(5), 1555–1562.

Williamson, J. (2001). Bayesian networks for logical reasoning. *AAAI Technical Report*, 136–143. Retrieved from `http://kar.kent.ac.uk/7396/`