Running head: A QUANTUM FRAMEWORK FOR PROBABILISTIC INFERENCE

A Quantum Probability Framework for Human Probabilistic Inference

Jennifer S. Trueblood

Vanderbilt University

James M. Yearsley

Vanderbilt University

Emmanuel M. Pothos

City University London

Corresponding Author:

Jennifer Trueblood

Department of Psychology

Vanderbilt University

PMB 407817

2301 Vanderbilt Place

Nashville, TN 37240-7817

email: jennifer.s.trueblood@vanderbilt.edu

Abstract

There is considerable variety in human inference (e.g., a doctor inferring the presence of a disease, a juror inferring the guilt of a defendant, or someone inferring future weight loss based on diet and exercise). As such, people display a wide range of behaviors when making inference judgments. Sometimes, people's judgments appear Bayesian (i.e., normative), but in other cases, judgments deviate from the normative prescription of classical probability theory. How can we combine both Bayesian and non-Bayesian influences in a principled way? We propose a unified explanation of human inference using quantum probability theory. In our approach, we postulate a hierarchy of mental representations, from 'fully' quantum to 'fully' classical, which could be adopted in different situations. In our hierarchy of models, moving from the lowest level to the highest involves changing assumptions about *compatibility* (i.e., how joint events are represented). Using results from three experiments, we show that our modeling approach explains five key phenomena in human inference including order effects, reciprocity (i.e., the inverse fallacy), memorylessness, violations of the Markov condition, and anti-discounting. As far as we are aware, no existing theory or model can explain all five phenomena. We also explore transitions in our hierarchy, examining how representations change from more quantum to more classical. We show that classical representations provide a better account of data as individuals gain familiarity with a task. We also show that representations vary between individuals, in a way that relates to a simple measure of cognitive style, the Cognitive Reflection Test.

Keywords: Human judgment, quantum probability theory, Bayes' rule, order effects, Markov condition

A Quantum Probability Framework for Human Probabilistic Inference

Everyday we face situations where we must make inferences about the world around us. For example, a doctor must determine the likelihood that a patient has a disease based on a set of symptoms. A juror must decide the probability that a defendant is guilty after hearing the cases made by the prosecution and defense. Or, maybe you want to judge the likelihood that you will weigh less next month if you start exercising more regularly and you improve your diet. In general, the inference problem involves judging the likelihood of some hypothesis (e.g., presence of a disease, guilt of a defendant, future weight loss) based on a series of evidence (e.g., medical symptoms, prosecution and defense cases, changes in your diet and exercise).

Bayesian inference is widely accepted as the normative approach to inference. However, decades of research in human judgment and decision-making have suggested that people's judgments often violate the rules of Bayesian inference and classical probability theory (Tversky & Kahneman, 1975). Despite this large literature, there is growing interest in which aspects of human judgment are consistent with normative prescriptions.

One can argue that the strongest empirical evidence for Bayesian principles in human inference comes from the domain of causal reasoning. Causal graphical models (CGMs) or models based on causal Bayes nets have been successful at explaining and predicting a wide range of behavior in causal inference. In a CGM, variables are represented as nodes and directed edges between nodes capture causal relations. These models represent causal relationships using Bayes' calculus (Kim & Pearl, 1983; Pearl, 1988) with additional assumptions for how interventions work and are considered to provide a normative account of causal judgment. CGMs have been shown to provide distinct predictions for causal inferences driven by observational, intervention-based, and counterfactual situations (Hagmayer, Sloman, Lagnado, & Waldmann, 2007). There are also various models built using this and similar frameworks that account for causal learning, specifically, reasoning based on learning through observation or intervention (intended to simulate experience based learning), or learning from statistical or contingency information.

In spite of the success of CGMs, some recent empirical studies report violations of the predictions of these models. For example, all Bayesian networks must satisfy a condition called the Markov property (this property is part of the definition of a Bayesian network; Russell & Norvig, 2003). This condition states that any node in a Bayesian network is conditionally independent of its nondescendents (e.g., noneffects), given its parents (e.g., direct causes). Informally, if we know about the causes of some event X, then the descendants of X may give us information about X, but the non-descendants cannot give us any more information about X. Recently, various studies (Rottman & Hastie, 2016, 2014; Park & Sloman, 2013; Rehder, 2014; Fernbach & Sloman, 2009; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008; Hagmayer & Waldmann, 2002) have provided evidence that people often violate the Markov condition when making causal inferences.

In another line of research, there has been an attempt to modify existing Bayesian models to explain away erroneous judgments (Costello, 2009; Costello & Watts, 2014). These models are interesting both from a philosophical point of view, because they may shed light on the principles underlying human reasoning, and also from a practical point of view, as an understanding of why we make judgment errors may inform strategies to improve decision making. Despite the promises of these approaches, they are currently limited in scope. The models developed are typically limited to a single judgment fallacy (such as the conjunction fallacy) and as far as we are aware, none of these approaches have been applied to inference directly (which ultimately involves judgments about a hypothesis given a sequence of observations).

We feel that there is a need for a comprehensive modeling framework that can account for a large range of empirical findings. It is clearly the case that both normative Bayesian and non-Bayesian principles are engaged in human inference. But how can we combine such disparate influences in a principled way? We propose a new framework for modeling human inference using quantum probability theory. In our approach, we postulate a hierarchy of mental representations, from ‘fully’ quantum to ‘fully’ classical, that could be adopted for different situations. Classical probability models represent one class of models in our hierarchy. The models in the hierarchy vary in the degree to which the representations are classical versus quantum, which is associated with the dimensionality of the model. High dimensional models (which involve more events represented according to classical probability theory) correspond to more complex mental representations whereas lower dimensional models represent simpler mental representations. The transition from one model to another in the hierarchy is effected through changing certain key assumptions, which in turn guides requirements for representation (this is the assumption of *compatibility*, which we will extensively consider shortly). The specific mental representation adopted for a problem might depend on a number of factors including task requirements, the complexity of the problem, experience, familiarity, and an individual's style of cognitive processing (we will focus on the latter two factors in this paper).

Quantum probability theory is the noncommutative analog of classical probability theory derived from quantum mechanics. As formalized by von Neumann (1932), quantum probability theory is a geometric approach to probability where events are represented as *subspaces* of a Hilbert space[1] (essentially a vector space). Note that we use the mathematical formalism of quantum theory without the associated physical meaning. In addition, we assume a fully classical neural substrate. By modeling events as subspaces rather than subsets, quantum probability theory entails a different logic than classical probability theory. The logic of quantum probability theory is the logic of subspaces which relaxes some of the assumptions of Boolean logic. In particular, quantum probability theory does not always have to obey the closure, commutative, and distributive properties.

Cognitive models based on quantum probability theory are computational level models (Marr, 1982), focusing on the principles and representations guiding human behavior (cf. Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). Quantum models are about what is computed (e.g., a choice preference or judgment of likelihood) and why the computation has the form it does (i.e., information from several variables needs to be combined; in some cases this leads to "impoverished" mental representations, which can be accounted for by quantum probabilities). Quantum cognition is presently best seen as a descriptive, formal theory for how we *do* reason under uncertainty, not how we *should* reason. A more detailed discussion of the explanatory scope of quantum models is provided in the General Discussion section.

Quantum cognitive models are typically employed for results that have so far been explained primarily on the basis of individual heuristics. Such models have been able to account for numerous findings in cognition and decision-making. These include violations of the sure thing principle (Pothos & Busemeyer, 2009), conjunction and disjunction fallacies (Busemeyer, Pothos, Franco, & Trueblood, 2011; Busemeyer, Wang, Pothos, & Trueblood, 2015), interference effects in perception (Conte, Khrennikov, Todarello, Federici, & Zbilut, 2009), violations of dynamic consistency (Busemeyer, Wang, & Trueblood, 2012), conceptual combinations (Aerts, Gabora, & Sozzo, 2013), interference effects of choice on confidence judgments (Kvam, Pleskac, Yu, & Busemeyer, 2015), the Ellsberg paradox (Haven & Sozzo, in press), and order effects in survey questions (Wang & Busemeyer, 2013). Research has also examined the mechanistic foundations of quantum probability models of cognition (Fuss & Navarro, 2013), showing how complex (classical) cognitive architectures can give rise to behavior that is best described in terms of quantum theory. Even though quantum cognitive models are descriptive, they typically allow novel predictions (arguably more so than heuristic accounts) and novel insights about the relevant psychological principles.

The difference between classical and quantum models is often phrased in terms of the way different events are represented by an individual, either as *compatible* or *incompatible* (we will explain these terms shortly). It is generally believed that experience with a particular situation, either from previous familiarity or acquired through learning, may allow events to be represented in a compatible way, whereas relatively novel situations are more likely to be represented in a incompatible way. In addition, quantum models are often used to explain similar phenomena as heuristics (Busemeyer et al., 2011), and so it seems plausible that incompatible representations of events, associated with quantum models, should be preferentially used for decisions executed spontaneously with little conscious deliberation.

Compatible events are ones that may be assigned a simultaneous truth value. Thus, if event X and event Y are compatible, their conjunction $X \wedge Y$ is well defined. The probabilities for compatible events obey the Kolmogorov axioms. Two immediate consequences are that for compatible events X and Y we have,

$$\begin{aligned} p(X \wedge Y) &= p(Y \wedge X), \\ p(X|Y) &= p(Y|X)\frac{p(X)}{p(Y)} \end{aligned} \tag{1}$$

Almost all events that we encounter in everyday life can in principle be represented in a compatible way. However doing so requires that decision makers have access to the joint probabilities of all of these events. This may be unfeasible, for example, from the point of view of memory capacity, since the number of probabilities grows exponentially with the number of events being considered. Equally, these probabilities might be difficult to compute, since joint probabilities correspond to subsets of the sample space. If it takes a finite number of previous experiences to learn the approximate measure of each subset, then the amount of experience required to compute a joint probability again grows exponentially with the number of events considered. For example, consider a situation where there are three binary events, X, Y, and Z, with values 1 or 2. The elementary events arise from the intersection of these three events and include events such as $X_1 \wedge Y_1 \wedge Z_2$. This sample space has $2^3 = 8$ elementary events. As the number of events increases, the dimensionality of the space required to represent all elementary events rapidly increases. For only six binary events, the dimension of the sample space is 64. This holds for both objective and subjective probabilities. For example, if X, Y and Z are the events 'rains tomorrow', 'pay raise', and 'job promotion', a compatible representation requires the existence of all joint events (such as 'it rains tomorrow and you get a promotion at your job, but do not get a pay raise') and the ability to assign coherent subjective probabilities to these events.

In contrast with compatible events, incompatible ones are those for which $X \wedge Y$ is undefined. Thus although the probabilities $p(X)$ and $p(Y)$ exist, the joint $p(X \wedge Y)$ may not. Typically one can define a modified version of conjunction with an explicit ordering, e.g. $X \wedge Y$ is taken to mean X *and then* Y for incompatible variables. This implies that $p(X \wedge Y) \neq p(Y \wedge X)$. Because joint events do not exist when events are incompatible, a lower dimensional space can be used to represent them. For example, if three binary events X, Y, and Z are all incompatible, then they can (minimally) be represented in a two dimensional space.

In quantum models, one can choose to model two events as either compatible or incompatible. If all events are chosen to be compatible one recovers a classical model, while if no two events are compatible (except for the trivial case of an event and its negation) then one has a maximally quantum model. If there are more than two possible events then there can be intermediate representations where some subset of events are compatible. Thus we should more accurately speak of a hierarchy of different representations, from fully quantum to fully classical. Note that we use the term 'hierarchy' in the colloquial sense because our models can be roughly ordered by dimensionality. However, the models in our 'hierarchy' are not nested.

We begin by describing how quantum probability theory gives rise to a hierarchy of mental representations for human inference. In the subsequent sections, we will discuss how our modeling approach explains a number of phenomena in human inference including order effects, reciprocity (i.e., the inverse fallacy), memorylessness, violations of the Markov condition, and anti-discounting. As far as we are aware, no existing theory or model can account for all five phenomena. Further, we provide some of the first empirical evidence for the co-occurrence of these effects. We also show that classical representations provide a better account of data as individuals gain familiarity with a task, and that mental representations (as captured by different models in our framework) can vary between individuals, in a way that relates to a simple measure of cognitive style, the Cognitive Reflection Test (Frederick, 2005).

A hierarchy of mental representations

Previously, many researchers have dealt with violations of the rules of classical probability theory, by elaborating rational models through the inclusion of extra assumptions (Costello, 2009; Costello & Watts, 2014) or by rejecting the applicability of classical probability theory wholesale and instead pursuing explanations based on heuristics (Tentori, Crupi, & Russo, 2013). For example, in the domain of causal inference, when behavior violates the rules of CGMs, such as the Markov condition, researchers have elaborated basic networks through the inclusion of hidden variables, which are causally related to all the observed variables and possibly arise from participants' general assumptions regarding the relevant stimuli. While these models often provide good accounts of data, the addition of hidden variables is often post hoc, included when a basic CGM fails to capture data. Also, it is sometimes difficult to reconcile hidden variable approaches with empirical data. For example, Rehder (2014) reported results that normative violations were equally likely in domains of economics, meteorology, and sociology as in an abstract (blank) domain. The assumption of hidden variables in a blank domain is unlikely, as such hidden variables are typically motivated from background knowledge considerations. Further, complex CGMs with multiple hidden variables are often difficult to conclusively test.

Rather than elaborating an existing classical model, we expand the range of probabilistic representations relevant in inference judgments. In all rational models, probabilistic inference follows the rules of classical probability theory. We relax this assumption and use quantum probability theory to build a hierarchy of models. Then, depending on the representation (and we will outline specific prescriptions for determining the appropriate representations), probabilistic calculations can be fully classical or demonstrate some of the peculiarities of quantum probability theory. That is, the approach can accommodate both Bayesian predictions and corresponding non-Bayesian deviations (in the specific way, allowed by quantum theory).

Our hierarchy of mental representations corresponds to different ways people might think about a particular problem. Levels in the hierarchy correspond to probabilistic models of different dimensionality with the highest dimensional model being fully classical and the lowest dimensional model being fully quantum. Mathematically, models with lower dimensionality involve more incompatible events than models with higher dimensionality. It is in this way that one could consider lower levels in the hierarchy as “more” quantum and higher levels in the hierarchy as “more” classical.

In this paper, we focus on an inference situation involving three binary variables, X, Y, and E, where X and Y are causes that independently influence an effect E. For example, E might be ‘future weight loss’ and X and Y represent diet and exercise. We decided to focus on this situation (also known as a common effect situation) because it is particularly common in everyday causal inference and has been extensively studied in the lab. It has also figured prominently in the development of models of causal reasoning (Griffiths & Tenenbaum, 2005; Cheng, 1997). Further, many of the effects we are interested in can be explored in this situation, without loss of generality. It also allows us a greater degree of comparability across different experiments, which is useful for detailed model analyses. Note that even though we concentrate our efforts on this particular situation, the modeling approach scales to a wider range of inference problems (including problems with more variables that have more than two outcomes).

In our hierarchy of mental representations, we can model the problem using two, four, or eight dimensional spaces. These different models correspond to different levels in the hierarchy with the 2-dimensional model describing a very simple representation of the problem (all events are incompatible so there are no joint events) and the 8-dimensional model describing the most complex representation of the problem (all events are compatible so all joint events exist). An advantage of this approach is that the relation between simpler and more complex representations can be defined very precisely. We describe the details of these different models in the following sections. An introduction to quantum probability theory is provided in the online supplementary material. Additional details about the model parameterizations can be found in Appendix A.

2-dimensional model

Consider the situation where someone is trying to judge his or her future weight loss. Let E represent the question “Will I weigh less next month?” with two possible answers $\{E_1, E_2\}$. That is, the answer can either be ‘true’ (E_1) or ‘false’ (E_2). In classical probability theory, we have a sample space S containing the elements E_1 and E_2. Since there are only two elements in our classical sample space S, the size of S is two.

Alternatively, with quantum probability theory, we can represent the very same event ‘weight loss’, with its possible outcomes E_1 and E_2, in a multidimensional Hilbert space. To do so, we replace the sample space S with a Hilbert space H where the elements E_1 and E_2 are associated with basis vectors of that space. In linear algebra, basis vectors are linearly independent vectors that span the space H. Since, in our example, there are only two basis vectors (associated with the elements E_1 and E_2), H is a 2-dimensional Hilbert space. (See online supplementary material for a more extensive introduction to quantum probability theory.)

There is a direct correspondence between the elements of a classical sample space and the basis vectors of a quantum model. Let us assume the classical sample space is the set $\{E_1, E_2\}$ and the corresponding Hilbert space has basis $\{|E_1\rangle, |E_2\rangle\}$. Here we are employing Dirac notation (also known as bra-ket notation) to represent the vectors. In Dirac notation, $|E_1\rangle$ is a column vector and $\langle E_1|$ is a row version of this vector. This notation is a convention borrowed from physics that is useful in simplifying algebraic expressions. We can choose a specific basis to represent vectors in H. For example, we can let the 2x1 column vector that has all zeros except for a one in the first row be a coordinate representation of the basis vector $|E_1\rangle$. Likewise, we can let the 2x1 column vector that has all zeros expect for a one in the second row be a coordinate representation of the basis vector $|E_2\rangle$:

$$|E_1\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |E_2\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2}$$

Thus, we can represent any elements of classical probability theory (in our example, E_1 and E_2) as basis vectors of a vector space.

Now, suppose the individual also considers diet and exercise. Let X represent the question "Did I follow my diet this month?" with possible answers $\{X_1, X_2\}$. Likewise let Y represent the question "Did I meet my exercise goals for the month?" with possible answers $\{Y_1, Y_2\}$. That is, the answers can either be 'true' (X_1 and Y_1) or 'false' (X_2 and Y_2). From the classical probability standpoint, adding events simply involves redefining the sample space S as the set containing the elements $E_1 \wedge X_1 \wedge Y_1$, $E_1 \wedge X_1 \wedge Y_2$, etc. However, in quantum probability theory, we must make a decision about the relationship between E, X, and Y. Two or more events can either be compatible or incompatible.

We might hypothesize that thinking about diet and exercise would influence thoughts about weight loss. That is, processing one event can interfere with processing the other events. This intuition is captured in the quantum model by using incompatible events. In the 2-dimensional model, we assume all three variables are incompatible. When two or more events are incompatible, we use a different basis for H to describe each event. As described earlier, we can represent the variable E in a 2-dimensional Hilbert space. If we assume all three variables are incompatible, we can represent X and Y as different bases for the same 2D space. Using Dirac notation, the pairs $\{|X_1\rangle, |X_2\rangle\}$, $\{|Y_1\rangle, |Y_2\rangle\}$, and $\{|E_1\rangle, |E_2\rangle\}$ provide three different bases of the space, corresponding to the variables X, Y, and E. In this model, all of the events correspond to simple rays. This is the reason the space is 2 dimensional.

In quantum probability theory, each event is associated with a projector. We define the projector for each basis vector as the outer product, such as the projector for the basis vector $|E_1\rangle$:

$$P_{E_1} = |E_1\rangle\langle E_1| = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{3}$$

This projector is simply a 2x2 matrix with zeros everywhere except for a one on the first diagonal. More generally, a projection operator is a mapping from the Hilbert space into a (typically smaller) subspace, which is idempotent (i.e., $P^2 = P$). The projector for the entire space, H, is the identity operator (e.g., a 2x2 matrix with zeros everywhere except ones on the diagonal).

Quantum probability theory also postulates the existence of an initial knowledge state ρ that represents the current state of the system. In particular, ρ is a matrix called a 'density operator' and can be thought of as describing the distribution of knowledge states of a group of heterogeneous participants. The probability of an event, such as E_1, given the initial knowledge state ρ is

$$p(E_1) = \mathrm{Tr}(P_{E_1}\rho) = \langle E_1|\,\rho\,|E_1\rangle \tag{4}$$

where Tr denotes the trace of a matrix (i.e., the sum of the elements on the main diagonal).

A critical part of quantum theory is defining the relationships between incompatible events. Consider the variables E and X. We can relate these two variables through a 'rotation'. Mathematically, we can obtain the X basis by applying a unitary transformation (i.e., rotation) to the E basis vectors. A unitary transformation is a matrix R that satisfies $R^\dagger R = I$ where I is the identity matrix and $R^\dagger$ is the conjugate transpose of R. The matrix R must be unitary to preserve lengths and inner products thus maintaining the properties of the Hilbert space which allow probability calculations. We chose to parameterize these 2D unitary transformations in the following way

$$R_j = \begin{pmatrix} \cos(\theta_j) & -\sin(\theta_j)e^{i\phi_j} \\ \sin(\theta_j)e^{-i\phi_j} & \cos(\theta_j) \end{pmatrix}. \tag{5}$$

Because there are three events, we need three unitary transformations to relate them, denoted R_E, R_X, and R_Y. We will take the initial state to be a diagonal matrix, $\rho = \mathrm{diag}(\rho, 1-\rho)$. The parameter ρ gives a measure of the heterogeneity of participants, if $\rho = 0$ or 1 participants are perfectly homogenous with respect to their representation of these events, values of ρ closer to 0.5 indicate greater participant heterogeneity. One of the ϕ_i parameters may be set to 0 without loss of generality (we will choose $\phi_E = 0$.) Thus we have six parameters in total for the 2D model,

$\{\rho, \theta_E, \theta_X, \phi_X, \theta_Y, \phi_Y\}$. Please see Appendix A for more details about the model parameterization.

In quantum probability theory, to calculate the conjunction of two incompatible events, we apply a sequence of projections. For example, if we wanted to calculate the probability of weight loss and dieting (the event $E_1 \wedge X_1$), we first have to decide the order of the projections. We can project our knowledge state ρ onto either the $|E_1\rangle$ basis vector or the $|X_1\rangle$ basis vector first. When two events are incompatible, the order of events matters. Specifically,

$$p(E_1 \wedge X_1) = Tr(P_{X_1} P_{E_1} \rho P_{E_1}) \neq Tr(P_{E_1} P_{X_1} \rho P_{X_1}) = p(X_1 \wedge E_1) \tag{6}$$

In the left hand side of the equation, we project first onto $|E_1\rangle$ and then onto $|X_1\rangle$. In the right hand side of the equation, we project first onto $|X_1\rangle$ and then onto $|E_1\rangle$. Thus, using incompatible events naturally gives rise to order effects. This is difficult to achieve in a classical probability model without building extra assumptions into the model.

One interesting feature of the 2D model is that because all the events are represented by projection operators onto one dimensional subspaces, various expressions for the probabilities simplify. One example is known as reciprocity,

$$\begin{aligned} p(X_1|Y_1) =& \frac{\mathrm{Tr}(P_{X_1} P_{Y_1} \rho P_{Y_1})}{\mathrm{Tr}(P_{Y_1} \rho)} = \frac{\langle X_1|Y_1\rangle \langle Y_1|\rho|Y_1\rangle \langle Y_1|X_1\rangle}{\langle Y_1|\rho|Y_1\rangle} \\ =& |\langle X_1|Y_1\rangle|^2 = \frac{\langle Y_1|X_1\rangle \langle X_1|\rho|X_1\rangle \langle X_1|Y_1\rangle}{\langle X_1|\rho|X_1\rangle} = p(Y_1|X_1) \end{aligned} \tag{7}$$

where the conditional probabilities are the same for two events, regardless of which event is the conditionalizing one. Another example is the memorylessness property,

$$p(E_1|Y_1, X_1) = \frac{\mathrm{Tr}(P_{E_1} P_{X_1} P_{Y_1} \rho P_{Y_1} P_{X_1})}{\mathrm{Tr}(P_{X_1} P_{Y_1} \rho P_{Y_1})} = |\langle E_1|X_1\rangle|^2 = \frac{\mathrm{Tr}(P_{E_1} P_{X_1} \rho P_{X_1})}{\mathrm{Tr}(P_{X_1} \rho)} = p(E_1|X_1) \tag{8}$$

where the probability of an event only depends on the most recent information given (in this example X_1). The above equations hold for all variables (e.g., $p(X_2|Y_1) = p(Y_1|X_2)$,

$p(X_1|Y_2) = p(Y_2|X_1)$, etc.). Also, note that we use the notation $p(E|Y,X)$ for the conditional probability of E given combined information about Y and X where information about event Y is presented before information about event X.

In sum, the 2D model is the maximally quantum model because all events are incompatible. Psychologically, it assumes that individuals consider one variable at a time and do not have mental representations of joint events. Judgments about multiple variables are performed by considering each variable sequentially.

2-dimensional POVM model

The 2D model discussed above makes the important assumption that the events in question are totally isolated from all other events. We do not know a priori whether this strong assumption will hold even in some cases. It seems reasonable that a participant's knowledge state at any moment in time will contain information about more than just $\{E,X,Y\}$. Thus, it makes sense to generalize the fully incompatible quantum model in a way that allows for 'weak' influences from other knowledge (Fodor, 1983). In this section, we describe the 2D POVM model, which is fully incompatible like the 2D model, but allows for weak interactions with other events. As we will see shortly, one important consequence of this model is that the "special" properties of reciprocity and memorylessness no longer hold. It is possible that an individual has an incompatible representation of all events, but does not display one (or both) of reciprocity and memorylessness. For example, Rehder and Burnett (2005) found that participants' judgments about the presence of an effect increased with the number of presented causes, thus violating memorylessness.

We start by assuming that a participant's complete knowledge space is large. The natural way to formulate such a model is by representing variables using higher dimensional subspaces while maintaining the incompatibility assumption. The price we pay for this increase in generality is that judgments are no longer represented by projection operators, but by the more general class of Positive Operator Valued Measures (POVMs; Nielsen & Chuang, 2000; Yearsley, in press;

Busch, Grabowski, & Lahti, 1995). POVMs are a generic way to consider measurements on states which are embedded in a larger space. As formulated in the Neumark Dilation Theorem (Busch et al., 1995), any POVM acting on a low dimensional space can be thought of as arising from a set of projective measurements on a higher dimensional space. In particular, the 2D POVM model represents a situation where the knowledge space is higher dimensional, because there are many other possible states in this space, which interact weakly with the three variables of interest $\{E, X, Y\}$.

Whenever variables interact, noise (or error) is introduced into the system. The degree of error is likely to depend on the particulars of the situation and so we leave it as a free parameter. Thus, we proceed by proposing a form for the POVM involving a single extra parameter ε, and then fit this parameter to the data. One way of interpreting ε is to note that it controls the extent to which an event and its negation are not orthogonal. As such it can be thought of as a measure of the number of other events that can either cause, or be caused by, both the event and its negation.

We will adopt a particularly simple type of POVM where the projection operators used to represent a measurement, for example

$$P_{E_1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad P_{E_2} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{9}$$

are replaced with the following ‘measurement operators’,

$$M_{E_1} = \begin{pmatrix} \sqrt{1-\varepsilon} & 0 \\ 0 & \sqrt{\varepsilon} \end{pmatrix}, \quad M_{E_2} = \begin{pmatrix} \sqrt{\varepsilon} & 0 \\ 0 & \sqrt{1-\varepsilon} \end{pmatrix} \tag{10}$$

These measurement operators are ‘complete’ in the sense that, $M_{E_1}M_{E_1}^{\dagger} + M_{E_2}M_{E_2}^{\dagger} = I$, but they are not orthogonal or idempotent. Note that we use a single ε for all variables. Thus we have seven parameters in total for the 2D POVM model, $\{\rho, \theta_E, \theta_X, \phi_X, \theta_Y, \phi_Y, \varepsilon\}$.

This happens to be a very straightforward example of a POVM in that the measurement operators may be written as,

$$M_{E_1} = \sqrt{1-\varepsilon}P_{E_1} + \sqrt{\varepsilon}P_{E_2} \tag{11}$$

and similarly for M_{E_2}. This form makes it obvious that the 2D POVM model reduces to the 2D model when $\varepsilon \to 0$, which will be important later.

Now we can work with these measurement operators to compute the various probabilities of interest, but there turns out to be a useful approximation that simplifies the algebra. In a typical psychology experiment we expect the value of ε to be of the order of $1-5\%$ (e.g., Yearsley & Pothos, 2016). This is because one effect of the POVM is to introduce apparently erroneous responses, i.e. the model will, with probability $\sim \varepsilon$, output 'false' when the answer is obviously 'true'. Such errors do happen in experiments, but for a relatively simple experimental set up such as the one we report in this paper the rate of such errors is expected to be low.

Under the assumption of small ε we can expand out all of our expressions for the predicted probabilities in powers of ε and keep only the lowest order terms. The lowest (non-zero) power of ε that appears in the probabilities is $\sqrt{\varepsilon}$, so we will keep only terms up to this order in what follows. We can therefore write,

$$M_{E_1} = P_{E_1} + \sqrt{\varepsilon}P_{E_2} + O(\varepsilon) \tag{12}$$

The first thing to note is that $M_{E_1}M_{E_1}^{\dagger} = P_{E_1} + O(\varepsilon)$, therefore the simplest probabilities, $p(E_1)$, $p(Y_1)$, and $p(X_1)$ etc. are the same in the 2D and 2D POVM models. Probabilities involving conjunctions and conditionals differ between the 2D and 2D POVM models. In particular, in contrast to the standard 2D model, conditionals in the 2D POVM are not symmetric under the interchange of events (e.g., $p(X_1|Y_1) \neq p(Y_1|X_1)$). This is how the 2D POVM model avoids reciprocity and memorylessness.

4-dimensional models

When some of the variables are compatible and others are not, we have a mixed classical / quantum model. In the case of three binary variables, these mixed models are all 4-dimensional. Similar to the 2D POVM model, reciprocity and memorylessness do not hold in 4D models (but for these higher dimensionality models, we do not consider POVM versions, so as not to increase their complexity too much). We consider two possibilities: (1) a model where the causes X and Y are compatible, but neither are compatible with the effect E, and (2) a model where X and E are compatible and Y and E are compatible, but the two causes X and Y are incompatible. There are other possible configurations of compatible and incompatible events. We focus on a common effect situation, which can be considered symmetric in the causes since both X and Y independently cause E. Thus other variants are unsatisfactory as they treat X and Y asymmetrically.

4D model with incompatible causes (4D$_{IC}$). In this model, it is assumed that individuals form mental representations for single cause and effect relationships, but do not think about multiple causal relationships simultaneously. This model posits that the two causes X and Y are each compatible with the effect (e.g., diet and exercise are both compatible with weight loss), but not with each other. We can think of this as saying people learn the causal relations $X \to E$ (e.g., diet causes weight loss) and $Y \to E$ (e.g., exercise causes weight loss) first, before learning the relationship between X and Y. An important consequence of this is that the variables E_1 and E_2 always look classical in this model. Although the space contains states which are indefinite with respect to the value of E (technically superposition states), the effects of this cannot be seen in the model predictions. We will use this below to reduce the number of parameters needed to describe the model.

Since X and E are compatible and are both binary variables, that means they form a set of 4D vectors that span the space. Likewise we can form another basis from Y and E. We therefore have the two bases,

$$\begin{pmatrix} X_1E_1 \\ X_2E_1 \\ X_1E_2 \\ X_2E_2 \end{pmatrix}, \quad \begin{pmatrix} Y_1E_1 \\ Y_2E_1 \\ Y_1E_2 \\ Y_2E_2 \end{pmatrix} \tag{13}$$

which are linked by a unitary transformation R. Now suppose we have an initial state ρ. The transformation $\rho \to R\rho R^\dagger$ should leave E unchanged (i.e. $\mathrm{Tr}(P_E\rho) = \mathrm{Tr}(P_E R\rho R^\dagger)$) so we conclude,

$$R = \begin{pmatrix} \cos(\theta_1) & -\sin(\theta_1)e^{i\phi_1} & 0 & 0 \\ \sin(\theta_1)e^{-i\phi_1} & \cos(\theta_1) & 0 & 0 \\ 0 & 0 & \cos(\theta_2) & -\sin(\theta_2)e^{i\phi_2} \\ 0 & 0 & \sin(\theta_2)e^{-i\phi_1} & \cos(\theta_2) \end{pmatrix} \tag{14}$$

where $\theta_1, \theta_2, \phi_1, \phi_2$ are real angles. Finally we need to specify the initial state. In general we have a 4×4 initial density matrix. However we noted above that because X and Y commute with E, we may take the initial state to be diagonal in the E_1 and E_2 basis. We can therefore write,

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} & 0 & 0 \\ \rho_{21} & \rho_{22} & 0 & 0 \\ 0 & 0 & \rho_{33} & \rho_{34} \\ 0 & 0 & \rho_{43} & \rho_{44} \end{pmatrix} \tag{15}$$

This would suffice to specify all the parameters in the $4\mathrm{D_{IC}}$ model. However it is useful to rewrite ρ, because it is difficult to ensure that a given choice of $\{\rho_{ij}\}$ leads to an allowable density matrix.

For this reason it is useful to write, $\rho = S\rho' S^\dagger$ where,

$$\rho' = \begin{pmatrix} \rho'_{11} & 0 & 0 & 0 \\ 0 & \rho'_{22} & 0 & 0 \\ 0 & 0 & \rho'_{33} & 0 \\ 0 & 0 & 0 & \rho'_{44} \end{pmatrix}, \quad S = \begin{pmatrix} \cos(\theta_a) & -\sin(\theta_a) & 0 & 0 \\ \sin(\theta_a) & \cos(\theta_a) & 0 & 0 \\ 0 & 0 & \cos(\theta_b) & -\sin(\theta_b) \\ 0 & 0 & \sin(\theta_b) & \cos(\theta_b) \end{pmatrix} \tag{16}$$

Here θ_a and θ_b are real angles. Comparing S with R we see S is not the most general unitary transformation of this form. We may restrict our attention to real S via an argument similar to that given in the 2D case (see Appendix A). The 4D$_{\mathrm{IC}}$ model thus contains 10 parameters, $\{\rho_{11}, \rho_{22}, \rho_{33}, \rho_{44}, \theta_1, \theta_2, \phi_1, \phi_1, \theta_a, \theta_b\}$. Note that the normalization constraint (i.e., $\rho_{11} + \rho_{22} + \rho_{33} + \rho_{44} = 1$) means that the degrees of freedom in the model is one less than the number of parameters (that is, the degrees of freedom is 9).

4D model with compatible causes (4D$_{CC}$). This model posits that the two causes X and Y are compatible with each other (e.g., diet and exercise are compatible), but neither is compatible with the effect E. Psychologically, this is reasonable since X and Y do not causally influence each other in the common effect situation. For example, it is easy to imagine a person that both diets and exercises. In our experiments (discussed below), participants made judgements about the casual relationships of features of novel animals such as African Lake Shrimp. In this case, one might imagine an exemplar (e.g., a particular shrimp) possessing both features X and Y simultaneously.

We can form two bases,

$$\begin{pmatrix} X_1Y_1 \\ X_1Y_2 \\ X_2Y_1 \\ X_2Y_2 \end{pmatrix}, \quad \begin{pmatrix} E_1\mathcal{E}_1 \\ E_1\mathcal{E}_2 \\ E_2\mathcal{E}_1 \\ E_2\mathcal{E}_2 \end{pmatrix} \tag{17}$$

Here $\mathcal{E}$ is a label to distinguish the different states having the same value of E. The subspaces corresponding to E_1 and E_2 are two dimensional, so they require two vectors to span them, which we label $|E_1, \mathcal{E}_1\rangle$ and $|E_1, \mathcal{E}_2\rangle$ etc. Since we only measure E and not the value of $\mathcal{E}$, it is therefore an unobservable parameter. This will be important below, as it will allow us to reduce the number of parameters needed to describe the model.

The two bases are linked by a unitary transformation, R. Since we noted that the value of $\mathcal{E}$ is unobservable when computing the probabilities for E_1 and E_2 we can choose the simple form,

$$R = \begin{pmatrix} \cos(\theta_1) & 0 & 0 & -\sin(\theta_1)e^{i\phi_1} \\ 0 & \cos(\theta_2) & -\sin(\theta_2)e^{i\phi_2} & 0 \\ 0 & \sin(\theta_2)e^{-i\phi_2} & \cos(\theta_2) & 0 \\ \sin(\theta_1)e^{-i\phi_1} & 0 & 0 & \cos(\theta_1) \end{pmatrix} \tag{18}$$

where $\theta_1, \theta_2, \phi_1, \phi_2$ are real angles. We can also choose the initial state ρ to be of the form,

$$\rho = \begin{pmatrix} \rho_{11} & 0 & 0 & \rho_{14} \\ 0 & \rho_{22} & \rho_{23} & 0 \\ 0 & \rho_{32} & \rho_{33} & 0 \\ \rho_{41} & 0 & 0 & \rho_{44} \end{pmatrix} \tag{19}$$

In a similar way to the $4D_{\text{IC}}$ case it is useful to express this as, $\rho = S\rho' S^\dagger$ where

$$\rho' = \begin{pmatrix} \rho'_{11} & 0 & 0 & 0 \\ 0 & \rho'_{22} & 0 & 0 \\ 0 & 0 & \rho'_{33} & 0 \\ 0 & 0 & 0 & \rho'_{44} \end{pmatrix}, \quad S = \begin{pmatrix} \cos(\theta_a) & 0 & 0 & -\sin(\theta_a) \\ 0 & \cos(\theta_b) & -\sin(\theta_b) & 0 \\ 0 & \sin(\theta_b) & \cos(\theta_b) & 0 \\ \sin(\theta_a) & 0 & 0 & \cos(\theta_a) \end{pmatrix} \tag{20}$$

where θ_a, θ_b are real angles, and the restriction to real S is allowed for a similar reason as for $4D_{\text{IC}}$. The $4D_{\text{CC}}$ model thus contains 10 parameters, $\{\rho_{11}, \rho_{22}, \rho_{33}, \rho_{44}, \theta_1, \theta_2, \phi_1, \phi_1, \theta_a, \theta_b\}$. Note that the normalization constraint means that the degrees of freedom in the model is one less than the number of parameters.

8-dimensional model

The 8D model is the classical probability model where all three variables X, Y, and E are compatible, implying individuals have mental representations of all joint events (e.g., diet, exercise, and weight loss are represented simultaneously). Since all events are compatible, it is possible to assign truth values to propositions such as $X_1 \wedge Y_1 \wedge E_1$, and so the space needs to contain vectors representing these events. Therefore our space is 8D and can be described by the following basis,

$$\begin{aligned} |X_1Y_1E_1\rangle &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T, \\ |X_1Y_1E_2\rangle &= \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T, \\ &\vdots \\ |X_2Y_2E_2\rangle &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}^T \end{aligned} \tag{21}$$

and projection operators,

$$\begin{aligned} P_{X_1} =& |X_1Y_1E_1\rangle\langle X_1Y_1E_1| + |X_1Y_1E_2\rangle\langle X_1Y_1E_2| \\ &+ |X_1Y_2E_1\rangle\langle X_1Y_2E_1| + |X_1Y_2E_2\rangle\langle X_1Y_2E_2| \\ =& \operatorname{diag}(1,1,1,1,0,0,0,0) \text{ etc.} \end{aligned} \tag{22}$$

The initial state may be a general density matrix, however it turns out that the probabilities we compute are sensitive only to the diagonal elements of ρ. Therefore we may take,

$$\rho = \operatorname{diag}(\rho_{11}, \rho_{22}, \ldots, \rho_{88}) \tag{23}$$

It is easy to compute the various probabilities of interest in terms of the ρ_{ii}. There are therefore 8 parameters in the classical model, $\{\rho_{11}, \rho_{22}, \ldots, \rho_{88}\}$. Note that the normalization constraint means that the degrees of freedom in the model is one less than the number of parameters. Also note that because all events are compatible, unitary transformations are not needed here.

The 8D model is entirely equivalent to a general classical probability model. More specific classical models can be represented within our framework by placing restrictions on the parameters of the model. A well motivated approach is parameterizing the model in accordance with power PC theory (or causal power theory for short, Cheng, 1997; Novick & Cheng, 2004). This is a classical probability model that links observable covariation to causal powers, by assessing the power of candidate causes to generate specific effects taking into account underlying base rates and independent alternate causes. Specifically, causal power theory provides a formal way to capture the intuitive idea that one variable can influence another by exerting power over it. Each cause i is associated with a power parameter w_i capturing the power of the cause to produce the effect. In particular, there are two power parameters, w_X and w_Y, for the two causes of E. Cheng (1997) also assumed there could be alternative causes for the effect which might be known or unknown. These alternative causes are also associated with a power parameter labeled w_a. By the axioms of classical probability theory and the independence of X and Y we can write the joint probabilities for the three features as $p(E_i, X_j, Y_k) = p(E_i|X_j, Y_k)p(X_j)p(Y_k)$ where i, j, and $k \in \{0, 1\}$. Causal power theory assumes the conditional probability of the effect given the causes is computed using a "noisy-or" equation:

$$p(E_1|X_j, Y_k) = 1 - (1 - w_X)^j (1 - w_Y)^k (1 - w_a). \tag{24}$$

Thus, five parameters are needed to define all eight possible joint probabilities: the three power parameters, w_X, w_Y, and w_a, and the prior probabilities of the causes, $p(X_1)$ and $p(Y_1)$. The eight joint probabilities can then be mapped directly to the diagonal elements of ρ.

In this paper, we consider both the general form of the 8D model and the parameterization due to causal power. In particular, we include the most general form because we are interested in whether people's judgments obey the axioms of classical probability theory (since if the more general model is shown inferior, this would apply to all specific instantiations too).

Model predictions

The models in our hierarchy of representations can explain a number of phenomena in human inference. In this paper, we focus on five key findings: order effects, reciprocity (i.e., the inverse fallacy), memorylessness, violations of the Markov condition, and anti-discounting. While there is extensive empirical evidence for most of these phenomena, very few studies have examined the co-occurrence of these effects. For example, there has been extensive research on the Markov condition and anti-discounting in causal inference (e.g., Rehder, 2014). However, order effects and reciprocity are rarely studied in this domain. Order effects and reciprocity have been mainly examined in non-causal inference problems. As far as we are aware, there is no existing research examining all five phenomena in the same paradigm. Further, there has been little research examining individual differences in these phenomena. In this paper, we examine two factors (familiarity and cognitive thinking style) that are related to the size of the effects. Below we describe each of the phenomena in detail, including the previous empirical evidence.

Order effects

Order effects are a hallmark of incompatible events because these effects show that events do not commute and must be evaluated sequentially. Mathematically, this is because when two events X and Y are incompatible, their projectors do not commute ($P_X P_Y \neq P_Y P_X$). On the other hand, compatible events obey the commutative property and do not naturally produce order effects. Thus, order effects provide a critical test between models using compatible events and those using incompatible events.

A wealth of past research has shown that order of information often plays a crucial role in determining final judgments (see Hogarth & Einhorn, 1992, for a review). Order effects arise in a number of different situations ranging from judging the likelihood of selecting balls from urns (Shanteau, 1970) to judging the guilt of a defendant in a mock trial (Furnham, 1986; Walker, Thibaut, & Andreoli, 1972). In general, for a sequence of information X followed by Y, individuals are asked to judge $p(H|X,Y)$ for some hypothesis H. An order effect occurs when final judgments depend on the sequence of information so that $p(H|X,Y) \neq p(H|Y,X)$.

While there has been extensive research on order effects in non-causal inference problems, there has been little work examining these effects in causal inference. Most past research on order effects in causality has examined these effects in the context of causal learning (Dennis & Ahn, 2001; Collins & Shanks, 2002; Abbott, Griffiths, et al., 2011). For example, many experiments focus on a pair of events (e.g., Collins & Shanks, 2002, examined the relationship between radiation and mutation) where participants learn the causal relationship between the two events over a sequence of trials. After the learning stage, participants are asked to make judgments of causal strength between the events. Order effects are observed when early trials in the learning sequence favor one relationship (e.g., a positive relationship between radiation and mutation) and later trials in the sequence favor the opposite relationship (e.g., a negative relationship between radiation and mutation). Specifically, Dennis and Ahn (2001) found a primacy effect where early information in the learning stage was weighted more heavily in causal strength judgments. On the other hand, Collins and Shanks (2002) found that when intermediate judgments were introduced (causal strength judgments after every 10 learning trials), participants showed recency effects, weighting later information more heavily in their judgments.

The focus of the present paper is on order effects in causal inference rather than causal learning. In particular, participants in our experiments are directly provided information about causal relationships and do not learn this information over trials as in Dennis and Ahn (2001); Collins and Shanks (2002); Abbott et al. (2011). We are interested in whether the order in which different causes are presented influences judgments about the likelihood of an effect (e.g., do judgments of future weight loss depend upon the presentation order of information about diet and exercise?). This is in contrast with studies of causal learning that examine whether presentation order during learning influences judgments of causal strength.

Preliminary empirical evidence for order effects in causal inference was obtained in Trueblood and Busemeyer (2012). Here we aim to test this prediction more comprehensively. In Trueblood and Busemeyer (2012), participants read different scenarios each involving a single effect and two causes where one of the causes was present and the other was absent. For example, in one scenario, participants were asked about the likelihood that sales of a popular caffeine free soda will increase next year (the effect) given the advertising budget for the soda remains the same (the absent cause) and the soda company lowers the price of the drink (the present cause). Participants reported the likelihood of the effect before reading either cause, after reading one of the causes, and again after reading the remaining cause. For a random half of the scenarios, subjects judged the present cause before the absent cause. For the remaining half of the scenarios, the subjects judged the absent cause before the present cause. The results showed a significant recency effect where individuals placed more importance on the final cause. Experiments 1 and 3 in this work provide more extensive tests of order effects in causal inference.

Reciprocity

Reciprocity refers to the situation where individuals judge the probability of one variable given another to be the same as the probability when the variables are reversed, e.g. $p(E|X) = p(X|E)$. This phenomenon is also related to the inverse fallacy (Koehler, 1996; Villejoubert & Mandel, 2002) where individuals equate posterior and likelihood probabilities. That is, for a hypothesis H and data D, individuals judge $p(H|D) = p(D|H)$ where $p(H|D)$ is the posterior probability and $p(D|H)$ is the likelihood. The inverse fallacy has mainly been studied in non-causal inference problems. For example, Kahneman and Tversky (1972) demonstrated the fallacy in their *taxicab problem*, where participants were asked to judge the probability that a cab had been in an accident given that it was blue instead of green. In this problem, most participants judged $p(H|D)$ as $p(D|H)$. In causal inference, there is some preliminary evidence for this phenomenon. For example, in medical reasoning, clinicians often exhibit the fallacy when judging the probability of a disease (the cause) based on a set of symptoms (the effects) (Meehl & Rosen, 1955; Hammerton, 1973; Liu, 1975; Eddy, 1982). Despite the evidence for the inverse fallacy in these studies, Krynski and Tenenbaum (2007) suggested that the phenomenon is limited, only occurring when both probabilities have roughly the same value. Note that in our experiments, we examine reciprocity under a broad range of conditions.

In quantum probability theory, the *law of reciprocity* (Peres, 1998) states that if two events X and E are represented by a single dimensional subspace (i.e., a ray), then $p(X|E)$ is exactly the same as $p(E|X)$. In our hierarchy of representations, reciprocity is predicted by the 2D model because events are represented by rays. Reciprocity is not predicted by the 2D POVM, 4D, or 8D models because events are represented by multi-dimensional subspaces in these models (e.g., events are planes in the 4D models and three dimensional subspaces in the 8D model.) Although, the 2D POVM model can display weak forms of reciprocity when ε is very small.

In a common effect situation, there are two distinct ways to test reciprocity. One way is to examine reciprocity between the effect and causes (e.g., $p(E|X)$ versus $p(X|E)$). The other way is to examine reciprocity between the two causes (e.g., $p(X|Y)$ versus $p(Y|X)$). We will call these different comparisons "cause-effect" reciprocity and "cause-cause" reciprocity. Experiment 1 tests for "cause-effect" reciprocity, Experiment 2 tests both types of reciprocity, and Experiment 3 tests for "cause-cause" reciprocity.

Memorylessness

In quantum probability theory, if three (or more) incompatible events are each represented by a single dimension, then conditional probabilities involving two (or more) given events (e.g.,

$p(E|Y,X)$) exhibit a memoryless property. That is, the probability of an event (E) only depends of the most recent information given (X). Earlier information (Y) does not factor into the probability. Therefore, memorylessness predicts equality among conditionals such as $p(E|Y,X)$ and $p(E|X)$. Similar to reciprocity, memorylessness is predicted by the 2D model and not the 2D POVM, 4D, or 8D models, because the 2D model is the only one where events are represented by rays. Note that for small ε, the 2D POVM model can display weak forms of memorylessness.

As far as we are aware, there have been no direct empirical studies of memorylessness. Rehder and Burnett (2005) found that participants' judgments about the presence of an effect increased with the number of presented causes, contrary to memorylessness. We examine the evidence for this phenomenon in all three experiments. We note that the prediction of memorylessness is a bold one and it is clearly not a general property of human inference (this point is trivially true, since memorylessess is not consistent with a fully classical model). However, it is intriguing to consider whether there might be at least some situations where memorylessness is observed.

Similar to reciprocity, there are different ways to examine memorylessness in the common effect situation. One way is to examine the probability of the effect conditioned on the causes (e.g., $p(E|X)$ versus $p(E|Y,X)$). Another way is to examine the probability of a cause conditioned on the effect and other cause (e.g., $p(X|E)$ versus $p(X|E,Y)$). We will call the first method "cause-cause" memorylessness (since the conditioning events are both causes) and the second method "cause-effect" memorylessness (since the conditioning events are the effect and one of the causes). Experiments 1 and 3 test "cause-cause" memorylessness. Experiment 2 tests "cause-effect" memorylessness.

Violations of the Markov condition

Violations of the Markov condition are specific to causal inference because they deal with violations of the assumptions of CGMs. The Markov condition stipulates that any node (i.e., event) in a CGM is conditionally independent of its nondescendents when its parents are known. For example, in the common effect situation, the Markov condition implies that the two causes X and Y are conditionally independent (note that these variables do not have parents). Thus, the presence or absence of one cause should not affect judgments of the other cause. For example, this implies that $p(X_1|Y_1) = p(X_1|Y_2)$ and similarly when X and Y are reversed. Rehder (2014) documented a number of situations where individuals violate the Markov condition including the common effect situation.

Independence of the events X and Y is equivalent to the conditions $p(X_1,Y_1) = p(X_1)p(Y_1)$, $p(X_1,Y_2) = p(X_1)p(Y_2)$, etc. In general, for any of the models to display independence for a pair of events, two conditions must hold. First X and Y must be compatible and second the initial state must factorize as $\rho = \rho_X \otimes \rho_Y$. This then guarantees that,

$$p(X_1,Y_1) = \mathrm{Tr}(P_{X_1 \wedge Y_1}\rho) = \mathrm{Tr}(P_{X_1}\rho_X)\mathrm{Tr}(P_{Y_1}\rho_Y) = p(X_1)p(Y_1) \qquad (25)$$

Since X and Y are incompatible in the 2D, 2D POVM and $4\mathrm{D}_{\mathrm{IC}}$ models, these models cannot display independence for these variables, except in some trivial cases. The 8D and $4\mathrm{D}_{\mathrm{CC}}$ models can allow for independence, but whether they display it or not depends on the choice of initial state. While the standard CGM for a common effect situation assumes X and Y are independent by definition, the 8D model is more general and does not have this same restriction. We test for violations of the Markov condition in Experiments 2 and 3.

Anti-discounting behavior

Similar to the violations of the Markov condition, anti-discounting is specific to causal inference. Discounting occurs in the common effect scenario when one cause, say X_1, casts doubt on the other cause, Y_1. Mathematically, discounting implies $p(Y_1|E_1,X_1) < p(Y_1|E_1)$. In many causal situations, discounting is the normatively correct way to judge events (Morris & Larrick, 1995). For example, it is normatively correct to judge $p(Y_1|E_1) > p(Y_1|E_1,X_1)$ because the presence of X_1 in the second conditional sufficiently explains the presence of the effect and so renders the other cause redundant (i.e., the presence of X_1 discounts the other cause). In the conditional $p(Y_1|E_1)$, the value of X is unknown thus increasing the chance that the effect was brought about by cause Y. However, Rehder (2014) found that many individuals judge the unknown cause Y as highly probable based on the presence of the alternative cause X. That is, $p(Y_1|E_1,X_1) > p(Y_1|E_1)$. Anti-discounting behavior can be attributed to a causal dependency between X and Y, which naturally arises in the 2D, 2D POVM and $4D_{IC}$ models. We test for anti-discounting in Experiment 2.

Other Non-normative Effects

The effects discussed above are the most relevant ones when studying inferences about causally related events, however they are not the only possible non-normative effects we might observe. In particular, for any model with two incompatible events A and B we expect to observe violations of the law of total probability, e.g. $p(A_1) \neq p(A_1|B_1)p(B_1) + p(A_1|B_2)p(B_2)$ (Pothos & Busemeyer, 2009). Therefore, for example, while the $4D_{CC}$ model may not display some of the non-normative behavioral properties we examine (i.e., Reciprocity, Memorylessness, Violations of the Markov condition, and Anti-Discounting), it is able to account for situations where $p(Y_1)$ is judged less likely than $p(Y_1|E_1)p(E_1)$, for example.

Differences in mental representations

Because the models in the present approach are all embedded in the same hierarchy, we can utilize knowledge about the formal relations between models to make predictions for when we expect particular models to best describe behavior. In particular, we expect that familiarity with a task will encourage individuals to adopt a more classical representation (e.g., 8D model). Classical representations are more complex (in the sense that they involve information about all joint probabilities), so it is reasonable to expect that they form after longer experience with a task (or effort, see shortly). Thus, we expect to see transitions in our hierarchy from more quantum to more classical with experience. We also hypothesize that cognitive thinking style might also influence the type of representation an individual adopts. For example, an individual that tends to make intuitive or spontaneous decisions might adopt a more quantum representation. Some quantum models can be interpreted as heuristics or simplistic reasoning (Busemeyer et al., 2011). Quantum representations are simpler, as all probabilities can be represented in a lower dimensionality space and knowledge about joint events need not be represented, but the price one pays for this simplicity is the requirement that joints are evaluated sequentially. On the other hand, we expect individuals that tend to make deliberative decisions will be better modeled using more classical representations. We test both of these predictions in Experiment 3. In this experiment, participants gain familiarity with the task through repetition. We also include a simple measure of cognitive style, the Cognitive Reflection Test (Frederick, 2005), to distinguish intuitive and deliberative thinking styles.

Summary of Model Predictions

Not all models in our framework can cover all of the effects mentioned above. While the 2D model can cover all phenomena, the 2D POVM and 4D models can only predict order effects, violations of the Markov condition, and anti-discounting. The 2D POVM and 4D models cannot predict reciprocity and memoryless effects (except when ε is very small in the 2D POVM model).

The 8D model cannot cover any of these phenomena, except violations of the Markov condition and anti-discounting under specific configurations of the initial state. We note that it is possible for Bayesian models to produce order effects. In our modeling framework, the 8D model can be considered an ‘exchangeable’ Bayesian model. That is, the joint probability of any set of variables is independent of the ordering of those variables. However, more complex Bayesian models can violate the exchangeability property and produce order effects. For example, consider a Bayesian model with two additional variables O_1 that X is presented before Y and O_2 that Y is presented before X. In this case, we obtain $p(E|X,Y,O_1) \neq p(E|X,Y,O_2)$. However, without specifying $p(E) \times p(O_i|E) \times p(X|E,O_i) \times p(Y|E,O_i,X)$, this approach simply redescribes the empirical results. Also note that the causal power model is a special case of the more general 8D model and can at best cover the same phenomena as this model.

Table 1 summarizes the five models in our framework and the phenomena that each model covers. Our claim that different models correspond to particular styles of thinking in inference is an idea that has a precedence in literature. For example, Rehder's (2014) theory of causal inference includes both normative and associative components. Our approach is similar, but our aim is to express all relevant influences (both normative and non-normative) within the same integrated, probabilistic framework.

Table 2 provides a summary of the three experiments. Because there are a large number of different probability questions that can be asked when working with three binary variables (especially when you consider all of the different possible ways to condition on the variables), we took the approach of testing different questions in different experiments. The first row of Table 2 lists the effects tested in each experiment. In Experiment 1, we examine six different models (2D, 2D POVM, $4D_{IC}$, $4D_{CC}$, 8D, and 8D causal power). The modeling results of Experiment 1 show that the 2D POVM and $4D_{CC}$ models outperform the other two quantum models. Thus, we drop the 2D and $4D_{IC}$ models from the analyses of Experiments 2 and 3. We also find that both the general 8D model and 8D causal power models perform about the same. Thus, we drop the more restricted 8D causal power model from the analyses of Experiments 2 and 3. The table also provides an overview of the types of analyses that we perform for each experiment.

Experiment 1

The first experiment examines predictions of our modeling framework that have previously received less attention in the causal inference literature: order effects, reciprocity, and memorylessness. This experiment (along with Experiments 2 and 3) uses a paradigm developed by Rehder and Hastie (2001) and Rehder (2003b, 2003a) to study causal inference with novel categories. We selected this paradigm because we wanted participants to reason using linguistic descriptions of events rather than using statistical information or learning contingencies through observation. There seems little doubt that at least in some cases causal knowledge is acquired in such a direct, linguistic way, as opposed to using statistical information or learning contingencies through observation. Additionally, there is evidence that people's judgments about causal systems often deviate from classical probability theory when tasks are presented using linguistic descriptions (Sloman & Fernbach, 2011; Trueblood & Busemeyer, 2012; Rehder, 2014). Thus, we focus our efforts in this domain. In our task, participants are given a linguistic description of a novel category (e.g., African Lake Shrimp) and asked to judge the likelihood that certain features cause others. Specifically, participants are given information about how two independent features can influence a third feature. The language used to describe the features and their relationships is purposely vague as many real life situations do not involve precise information. Note that while our experiments use a paradigm similar to Rehder and Hastie (2001) and Rehder (2003b, 2003a), those previous studies did not examine the three effects of interest (order effects, reciprocity, and memorylessness). Thus, the aim of this experiment is part replication, and part examining these new effects. The data and models for all three experiments are available on the Open Science Framework at https://osf.io/4chu6/.

Methods

58 undergraduate students from a US university participated in the experiment online at a time of their choosing for course credit. Here and elsewhere, the sample size was determined a priori, broadly following Rehder (2014). Participants were randomly assigned to one of two novel animal categories (either African Lake Shrimp or Kehoe Ants). At the start of the experiment, they were told that biologists recently discovered a new type of animal (i.e., shrimp or ant) and that identical animals could be found in several different locations (e.g., identical shrimp can be found in all nine African Great Lakes). Participants then learned that each animal had three binary

features (X, Y, and E) where two of the features (X and Y) causally influenced the third (E), forming a common effect network. For example, in the African Lake Shrimp category, X_1 = high amount of ACh neurotransmitter (X_2 = low amount of ACh), Y_1 = accelerated sleep cycle (Y_2 = normal sleep cycle), and E_1 = high body weight (E_2 = low body weight). Participants were given information about the typicality of feature values. For example, they were told that "Most shrimp have a high amount of ACh whereas a few have a low amount of ACh". In both categories, most animals had feature X_1, a few had feature X_2, a few had feature Y_1, and most had feature Y_2. Also, half of the animals had feature E_1 and half had feature E_2. Participants were also given the causal relationships between features. These relationships were described as one feature causing another. In both categories, X_1 and Y_1 were described as causing E_1. Likewise, both X_2 and Y_2 were described as causing E_2. Participants were also told that there were no known relationships between X and Y. Details of the stimuli are given in Appendix B.

Participants first studied the three features and the typicality of their values. After studying this information, they took a multiple-choice test with six questions that tested them on this knowledge. Participants were required to answer each question correctly before moving on to the next one. Next, they studied the two causal relationships and took another multiple-choice test with eight questions testing them on this new knowledge. As before, participants were required to answer each question correctly before moving on to the next one. Finally, participants were asked to take a few minutes to review the features and relationships one more time. After they finished reviewing this information, they completed a third multiple-choice test with 10 questions. In this final test, participants were only given one opportunity to answer each question. Their score on this test was used to gauge how well they learned the features and causal relationships.

After completing the learning stage, participants completed two blocks of trials (i.e., two within-subject order conditions denoted by BX and BY) where they were asked to make decisions about the value of different features. Each block contained 13 questions where participants were asked to select the value of a particular feature (see Table 3). At the start of each question, participants were told that a biologist caught a new animal (either shrimp or ant) in a particular location (e.g., Lake Victoria) and were queried about one of the features of that animal. For example, in the African Lake Shrimp category, they might be asked ”What type of body weight do you think this shrimp has?” (question E in the table). Participants were given three response options: feature value 1, feature value 2, or equally likely to be feature value 1 or 2. For example, in the question about body weight, the response options were 1) a low body weight, 2) a high body weight, and 3) equally likely to be low or high.

Some of the questions required participants to make a sequence of decisions about a feature value (e.g., E) as they learned new information about the other features (e.g., X and Y). For example, they might be asked about the body weight of a shrimp given lab tests that showed the shrimp had a high amount of ACh neurotransmitter (i.e., $E|X_1$). Participants might then be asked to reevaluate body weight based on additional lab tests that showed the shrimp also had a normal sleep cycle (e.g., $E|X_1,Y_2$). Note that information about the value of the first feature (e.g., X_1) remained on the computer screen when new information about the second feature (e.g., Y_2) was presented. Thus, participants had access to all feature information during their final choice, which makes it less likely (if not impossible) that any observed effects result from memory failures.

Participants were randomly assigned to start with either the BX or BY block. In the BX (BY) block, information about feature X (Y) was always presented before information about feature Y (X) in sequences involving both features. This helped reduce the influence of memory on future decisions about reverse orderings. At the start of each block, participants were given a new location for the animals in that block (e.g., a different lake for the African Lake Shrimp). Changing the location of the animals between blocks helped delineate the blocks and obscure the repetitive nature of the questions. Participants did not receive any feedback about their choices.

Behavioral Results

We use Bayesian statistics for all analyses in this paper. All tests were implemented using the open source software package JASP (JASP Team, 2016). For each test, we report the Bayes factor (BF), which is a ratio quantifying the evidence in the data favoring one hypothesis relative to another. In particular, we report BF_{10}, which is the evidence for the alternative hypothesis relative to the null hypotheses. When $\text{BF}_{10} < 1$, there is evidence for the null hypothesis. When $\text{BF}_{10} > 1$, there is evidence for the alternative hypothesis. The larger BF_{10}, the more evidence there is in favor of the alternative hypothesis. While Bayes factors are directly interpretable, labels for the strength of the Bayes factor have been proposed. In particular, BF greater (less) than 1, 3 (1/3), 10 (1/10), 30 (1/30) and 100 (1/100) are considered 'Anecdotal', 'Moderate', 'Strong', 'Very Strong' and 'Extreme' evidence respectively (Kass & Raftery, 1995). We also note the coarseness of the response scale (only three response options). For the analyses and modeling below, we only consider group level results. Experiment 2 uses a different response scale and allows for individual level analyses.

All participants were included in the analyses. The average score on the 10 question multiple choice test was 9.41 indicating most participants correctly learned the feature values and causal relationships during the first part of the experiment. For analyses of the choice data, we calculated a *choice score* for each participant in a similar way to Rehder (2014) by assigning the following values to the three response options: feature value 1 = 1, feature value 2 = 0, and equally likely = 0.5[2]. Note that there were no differences between choices in the two different animal categories (BF_{10} = 0.105) and so responses were collapsed for the following analyses. The mean choice scores along with standard deviations for each question are given in Table 3.

Order effects were assessed by comparing pairs of questions such as $E|X_1,Y_2$ and $E|Y_2,X_1$. There are four possible comparisons that can be made by pairing the questions in block BX with the corresponding questions in block BY (see Table 3). Bayesian paired samples t-tests were conducted on the four pairs and the results are reported at the top of Table 4. Choices where both feature values matched (i.e., both X and Y equal to 1 or both equal to 2), showed no evidence for order effects. The lack of order effects for these questions could simply be due to floor and ceiling effects. As seen in Table 3, the mean choice scores for conditionals with two matching causes are either very close to 1 (when both X and Y equal 1) or very close to 0 (when both X and Y equal 2). On the other hand, there was 'very strong' to 'extreme evidence' for order effects when the feature values are mismatched. These order effects are easily seen in the mean choice scores reported in Table 3. The mean choice score for $E|X_1,Y_2$ was 0.42 as compared to 0.68 for $E|Y_2,X_1$. Likewise, the mean choice score for $E|X_2,Y_1$ was 0.63 as compared to 0.37 for $E|Y_1,X_2$. These results show that participants place more weight on recent information, demonstrating a recency effect.

Reciprocity or the inverse fallacy (Koehler, 1996; Villejoubert & Mandel, 2002) was examined by comparing pairs of questions such as $X|E_1$ and $E|X_1$. As a reminder, there are two distinct ways to test for reciprocity. One way is to examine reciprocity between the effect and causes, called "cause-effect" reciprocity (e.g., $X|E_1$ versus $E|X_1$). The other way is to examine reciprocity between the two causes, called "cause-cause" reciprocity (e.g., $X|Y_1$ versus $Y|X_1$). In this experiment, we only examined "cause-effect" reciprocity. In Experiments 2 and 3, we examine "cause-cause" reciprocity.

There are four possible comparisons that can be made by pairing the questions from both blocks. Bayesian paired samples t-tests were conducted on the four pairs and the results are reported in the middle of Table 4. Reciprocity occurs when the probability of one feature given another is the same as the probability when the features are reversed, e.g. $p(E_1|X_1) = p(X_1|E_1)$. In other words, reciprocity is an invariance and evidence for the effect is seen as evidence for the null hypothesis. Thus, when $\text{BF}_{10} < 1$, there is evidence for reciprocity. It is perhaps easier to evaluate the strength of evidence for reciprocity if we rewrite the Bayes Factor so that evidence for the null hypothesis is in the numerator (i.e., BF_{01}). This shows the effect ranges from BF_{01} = 0.91 ('anecdotal' evidence for the alternative hypothesis) in the comparison of $Y|E_1$ and $E|Y_1$ to BF_{01} = 6.71 ('moderate' evidence for the null hypothesis) in the comparison of $Y|E_2$ and $E|Y_2$.

Memorylessness occurs when the probability of a feature only depends on the most recent information given, e.g. $p(E_1|X_1) = p(E_1|Y_1, X_1)$ since X_1 is the most recent given information. Similar to reciprocity, there are different ways to examine memorylessness. One way is to examine the probability of the effect conditioned on the causes, called "cause-cause" memorylessness (e.g., $E_1|X_1$ versus $E_1|Y_1, X_1$). Another way is to examine the probability of a cause conditioned on the effect and other cause, called "cause-effect" memorylessness (e.g., $X_1|E_1$ versus $X_1|E_1, Y_1$). This experiment examines "cause-cause" memorylessness. Experiment 2 examines "cause-effect" memorylessness.

There are eight possible comparisons that can test for this property. Bayesian paired samples t-tests were conducted on all eight pairs and the results are reported at the bottom of Table 4. Similar to reciprocity, memorylessness is an invariance and evidence for the effect is seen as evidence for the null hypothesis (i.e., when $\text{BF}_{10} < 1$). The evidence for memorylessness is mixed. When the feature values of the causes match (i.e., both X and Y equal to 1 or both equal to 2), there is evidence for memorylessness. However, when the feature values of the causes are mismatched, there is strong evidence against memorylessness. In the case where feature values match, the result could simply be due to floor and ceiling effects because the mean choice scores in these questions are either very close to 1 or 0 (see Table 3).

Modeling Results

The behavioral results of Experiment 1 show evidence for order effects, reciprocity, and memorylessness (although evidence for the latter two effects is mixed). This recommends our modeling approach, which encompasses representations that can account for these effects. In this section, we explore this further by comparing six different models, ranging from fully quantum (all events are incompatible) to fully classical (no incompatible events).

General Modeling Procedures. Model fitting for all experiments was done using a Bayesian analysis carried out via the program Just Another Gibbs Sampler (JAGS; Plummer et al., 2003).

Three Markov chain Monte Carlo (MCMC) chains were used with 50,000 samples and a burn-in of 5000 samples. Chain convergence was assessed using the $\hat{R}$ statistic, and all chains had good convergence behavior.

Unless otherwise noted, the priors for all angle variables were taken to be $\frac{\pi}{2} \times$ Beta(2,2). For the general 8D model and two 4D models the priors for the ρ_{ii} variables were taken to be uniform in the interval $[0,1]$ and then normalized to ensure $\sum_i \rho_{ii} = 1$. For the causal power parameterization of the 8D model, the priors for w_X, w_Y, w_a, $p(X_1)$ and $p(Y_1)$ were uniform in the interval $[0,1]$. For the 2D and 2D POVM models it is more useful to set the priors to be asymmetric, as this helps convergence. The reason for this is that the quantum models are invariant under certain transformations of the variables; restricting the range of the ρ variable helps to avoid different chains converging on apparently different parameter sets which are in fact equivalent. The 2D model is particularly prone to this, and so we set the prior for ρ to be uniform in the range $[.5,1]$. The 2D POVM is less sensitive in this regard, and a prior for ρ taken to be uniform in the interval $[.2,1]$ produced good convergence behavior. For the 2D POVM model the prior for ε was taken to be uniform in the interval $[0,.05]$, which is based on empirical fits in previous work (Yearsley & Pothos, 2016).

For model comparisons, we used JAGS to compute the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). The DIC is a generalization of the BIC, with smaller values indicating a better fit. The DIC includes a component related to the goodness of fit as well as a component related to model complexity (technically, the effective number of parameters). Thus the DIC balances accuracy with parsimony. Note that the DIC is not directly interpretable, however differences between DIC values for models fit to the same data set can be interpreted. A difference in the DIC of 10 or more is usually taken to indicate a strong advantage in fit.

Results. We fit the models to the mean choice scores across participants. There were a total of 26 questions in the experiment. Because some questions were repeated, there were 19 unique questions. We fit the models to these 19 questions. Because our models output probabilities rather than choices, we transformed each predicted probability by a softmax function (similar to Rehder, 2014) to simulate the fact that each participant is forced to choose between the three alternatives rather than outputting a probability judgment. Specifically, we assumed that individuals represent probabilities as log odds and that selecting option x follows the softmax rule:

$$\text{choice}(x) = \frac{exp(\frac{logit(p_x)}{\tau})}{exp(\frac{logit(p_x)}{\tau}) + exp(\frac{logit(1-p_x)}{\tau})}. \tag{26}$$

where p_x is the predicted probability from the model and τ is a "temperature" parameter that controls the extremity of the responses. We then assumed that mean choice scores followed a beta distribution using the outputs of the softmax:

$$\text{choice score} \sim Beta(\lambda \times \text{choice}(x), \lambda(1 - \text{choice}(x))) \tag{27}$$

where the parameter λ controls the variance. The prior for τ was taken to be $N(0.1, 0.01)$ and the prior for λ was uniform in the range $[2, 100]$. Note that the addition of λ and τ adds two parameters to the parameter counts given in Table 1.

The DIC values for all six models are the following: 2D = -25.20, 2D POVM = -38.95, $4D_{IC}$ = -4.03, $4D_{CC}$= -42.55, general 8D= -25.32, causal power 8D = -25.88. The $4D_{CC}$ and 2D POVM models are clearly superior to the other quantum models and crucially they are superior to both versions of the 8D model. This confirms the behavioral results showing that this data contains marked non-classical effects. While the DIC is useful for model comparisons, it does not provide much information on how well a model describes the data. The model with the lowest DIC could still provide a terrible fit to observed data. To examine the fits of the models, we plotted the predictions from the six models against the data in Figure 1. The figure shows the average choice scores (red circles) for all questions plotted against the posterior distributions from the models. The size of the squares is proportional to the posterior mass. For the $4D_{CC}$ and 2D POVM models, the majority of choice scores fall within the area of the posterior distributions with the greatest mass, indicating that these models provide good fits to the data. The mean and 95% highest density intervals of the parameter estimates for each model are given in Appendix C[3].

Two features of these fits are worth particular note. First, the simplest 2D quantum model does not perform significantly better than either 8D model. This is interesting because it shows that while there may be quantum behavior here, it is more complex than one might have expected. Looking at the fit, it is clear that the 2D model fails to fit the data well because the memoryless property is not always obeyed in the data. This makes the value of the 2D POVM model clearer as it breaks the link between order effects and memorylessness, allowing it to better fit the data. Note that the mean value of ε is about 3.2%, meaning that the POVM elements fail to be orthogonal by a relatively small amount, nevertheless this is enough to improve the fit by a significant degree. Second, the worst performing model is the $4D_{IC}$ model. This model predicts that individual cause-effect relationships should look classical, which is a strong requirement.

We also performed a Bayesian hierarchical fitting of the models. Hierarchical Bayesian methods use both group-level and individual data for model fitting. Thus, models are fit to all of the data rather than simply group averages. The results of the hierarchical fitting were the same as the group-level results reported above. The $4D_{CC}$ and 2D POVM models had the best DIC values. The details and results of the hierarchical fitting are provided in the online supplementary material.

Validation of Model Fits. We also wanted to examine the issue of model flexibility and overfitting. For Experiment 1, we used 19 data points for the modeling. The number of parameters in the models ranged from 7 parameters in the causal power 8D model to 12 parameters in the $4D_{IC}$ and $4D_{CC}$ models (although the degrees of freedom for the 4D models is one less than the number of parameters due to the normalization constraint; see also Appendix C for a list of the parameters for each model). Because the number of data points to model parameters is small, this raises concerns about model flexibility and overfitting. It is worth noting that for the 2D and $4D_{IC}$ models there is large misfitting, as seen in the Figure 1. These two models have roughly the same number of parameters as the other models (in fact, the $4D_{IC}$ model has the largest number of parameters), yet they cannot account for the data. Thus, the good fits by the 2D POVM and $4D_{CC}$ models are probably not simply due to overfitting. Rather, the improved fits are most likely due to structural differences between the models. All of the models we examine have unique structural properties (e.g., commutativity in the general and causal power 8D models). Thus, there exist data patterns that each model cannot account for (e.g., the 8D models cannot account for order effects). This naturally places restrictions on model flexibility (see Table 1 for more details about the effects each model predicts).

To further examine the issue of overfitting, we used cross validation. We did this by randomly dividing the participants into two groups. We fit the models to the data for one group and then made predictions for the other group. We evaluated model performance for the second group using the mean squared error (MSE) between the model predictions (based on the mean of the posterior predictive) and the data. The results of the cross validation are reported in Table 5. We report the DIC value for the fits to the first group and the MSE for the second group. The results show that the best fitting models, the 2D POVM and $4D_{CC}$ models, also had the lowest MSE. This provides strong evidence that the good performance of these models is not due to overfitting.

Conclusions

Experiment 1 considers three phenomena (order effects, “cause-effect” reciprocity , and “cause-cause” memorylessness) that have traditionally received less attention in the causal inference literature. Overall, the behavioral and modeling results of this experiment paint a mixed picture. It is clear that there is at least some non-classical behavior, since models utilizing incompatible representations provide good fits. In particular, both 8D models are clearly outperformed by the $4D_{CC}$ and 2D POVM models. However the best fitting model (albeit by an small margin) predicts no order effects, when these are clearly present in the behavioral analysis in some conditions. Equally the success of the 2D POVM model over the simple 2D one suggests that not all quantum effects are present to the same degree.

It is not possible to separate the $4D_{CC}$ and 2D POVM models, or the 8D and 2D models, on the basis of this data set. It is clear, however, that the $4D_{IC}$ model performs significantly worse than the other models. In the interests of brevity and clarity we will therefore exclude this model from further analyses. Also, the 2D POVM model performs significantly better than the 2D model, and this is an interesting case because the 2D POVM reduces to the 2D model when $\varepsilon \rightarrow 0$ (the other models in the hierarchy do not have this simple property.) We can therefore drop the 2D model from further analyses; if the best fit parameter for ε in the 2D POVM model is extremely close to 0, we effectively have the simple 2D model. We also drop the casual power version of the 8D model since this model is a special case of the general one and model comparisons showed almost no difference between the two.

Given the evidence for a mixture of representations in Experiment 1, one reasonable hypothesis is that there are individual differences in the representations participants form. It could also help to explain why some quantum effects appear stronger than others - if not all models give rise to all quantum effects, and the representations used vary between participants, we would expect to see stronger evidence for those effects which are generic (e.g. order effects) and weaker evidence for those effects which are model specific (e.g. memorylessness).

Experiment 2

This experiment examines individual differences in reciprocity, memorylessness, violations of the Markov condition, and anti-discounting. We did not include questions about order effects in this experiment. Since the order effects in Experiment 1 were quite strong (for the mismatching causes), we wanted to focus on the other effects in this experiment. In particular, the results of Experiment 1 were inconclusive with regard to reciprocity and memorylessness. There was positive evidence for reciprocity, but this evidence was not strong. The evidence for memorylessness was mixed, some comparisons showed evidence for the effect, but others showed strong evidence against the effect. In the cases where there was evidence for memorylessness, the effect was confounded by floor and ceiling effects. The modeling results from Experiment 1 were also inconclusive. While the 2D POVM and $4D_{CC}$ models outperformed the 2D, $4D_{IC}$, and 8D models in terms of DIC, the two models were difficult to distinguish from each other. Taken together, these results suggest that there could be individual differences in the representations that people used in our task. Because we used a three choice response mode in Experiment 1, it is difficult to explore this possibility with that data set. Thus, we conducted a new experiment where participants provided probability judgments rather than choices, allowing us to model individuals rather than average data. This, together with a different choice for causal questions (corresponding to the shift in the emphasis for the effects examined), are the only differences between Experiments 1 and 2.

Methods

58 undergraduate students from a US university participated in the experiment online for course credit. The instructions, stimuli, and procedures were similar to Experiment 1. As before, this experiment used a within subjects design. However, participants provided probability judgments rather than choices in this experiment. Further, there was only a single novel category (i.e., the African Lake Shrimp) rather than the two categories used in the first experiment. At the start of the experiment, participants were told that shrimp could be found in three different lakes in Africa, coded as LakeA, LakeB, and LakeC. The features of the shrimp (i.e., ACh neurotransmitter, sleep cycle, and body weight) and the relationships between features were described exactly as in Experiment 1. The learning stage was identical to Experiment 1.

After completing the learning stage, participants answered 14 conditional judgment questions (see Table 6). Participants were asked to enter their likelihood judgments about specific features (e.g., sleep cycle or body weight) as numbers between 0 and 100 in a text box after reading each question. They were told that a judgment of 0 implied that they were certain the shrimp did not have the feature, a response of 50 implied that the shrimp was equally likely to have the feature or not, and a response of 100 implied that they were certain the shrimp did have the feature. On each trial, there was a scale from 0-100 reminding participants of this information. Participants judged all questions twice in two different randomized blocks. Participants did not receive feedback about their judgments.

Similar to Experiment 1, some questions involved participants making a sequence of two judgments. For example, participants might first read "A shrimp is caught in LakeA. After lab testing, you learn that the shrimp has a high body weight. Given this information, how likely is it that this shrimp has a high quantity of ACh neurotransmitter?" After providing a judgment, participants would then read "After further observation in the lab, you also learn that the shrimp has an accelerated sleep cycle. Given this new information, how likely is it that this shrimp has a high quantity of ACh neurotransmitter?" Similar to Experiment 1, the first question remained on the screen during the presentation of the second question, which was positioned directly below the first. Thus, participants had all relevant feature information available on the screen during the second judgment.

Behavioral Results

All participants were included in the analyses. The average score on the 10 question multiple choice test was 7.8 indicating most participants correctly learned the feature values and causal relationships during the learning stage of the experiment. For each participant, we first averaged the two judgments for each question. Mean judgments along with standard deviations for each question are given in Table 6.

Reciprocity was tested using three different pairs of judgments. Two pairs tested "cause-effect" reciprocity and one pair tested "cause-cause" reciprocity. Bayesian paired samples t-tests were conducted on the three pairs and the results are reported at the top of Table 7. As before, when $BF_{10} < 1$, there is evidence for reciprocity. If we rewrite the Bayes Factor so that evidence for the null hypothesis is in the numerator (i.e., BF_{01}), this shows the effect ranges from $\mathrm{BF}_{01} = 3.54$ in the comparison of $Y_1|E_1$ and $E_1|Y_1$ to $\mathrm{BF}_{01} = 6.38$ in the comparison of $X_1|E_1$ and $E_1|X_1$. This range of Bayes Factors is similar to that found in Experiment 1. In particular, the two "cause-effect" reciprocity pairs used in both experiments (i.e., $X|E_1$ versus $E|X_1$ and $Y|E_1$ versus $E|Y_1$) have similar Bayes Factors. For the comparison of $X|E_1$ and $E|X_1$, $\mathrm{BF}_{01} = 4.98$ in Experiment 1 and $\mathrm{BF}_{01} = 6.38$ in Experiment 2. This is 'moderate' evidence for reciprocity. For the comparison of $Y|E_1$ and $E|Y_1$, $\mathrm{BF}_{01} = 0.91$ in Experiment 1 and $\mathrm{BF}_{01} = 3.54$ in Experiment 2. This is 'anecdotal' to 'moderate' evidence.

The continuous response scale in this experiment further allows us to examine the extent to which reciprocity is an artifact of averaging or not. The top left panel in Figure 2 shows a scatter plot with the three different pairs of judgments for reciprocity. Each point represents an individual-level judgment. As can be seen in the figure, most of the data falls around the 45 degree line of identity, suggesting reciprocity occurs at the individual-level as well as the group-level. However, there are clearly large individual differences in the figure. We will address the sources of these individual differences using our modeling approach.

Memorylessness was tested using four different pairs of judgments. This experiment examines "cause-effect" memorylessness where the conditioning events are the effect and one of the causes (e.g., $X_1|Y_2$ versus $X_1|E_1,Y_2$). Bayesian paired samples t-tests were conducted on all four pairs and the results are reported in the middle of Table 7. Similar to reciprocity, when $\mathrm{BF}_{10} < 1$, there is evidence for memorylessness. Rewriting the Bayes Factor shows the effect ranges from $\mathrm{BF}_{01} = 1.67$ in the comparison of $X_1|Y_2$ and $X_1|E_1,Y_2$ to $\mathrm{BF}_{01} = 5.68$ in the comparison of $Y_1|X_2$ and $Y_1|E_1,X_2$. Thus evidence for memorylessness is 'anecdotal' to 'moderate'. Note that Experiment 1 examined "cause-cause" memorylessness, thus it is difficult to directly compare the results of this experiment to the results of Experiment 1. However, the Bayes Factors are in the same range as comparisons where the feature values of the causes were matched in Experiment 1. The top right panel in Figure 2 shows a scatter plot with the four different pairs of judgments for memorylessness. As can be seen in the figure, most of the data falls around the 45 degree line of identity, suggesting memorylessness occurs at the individual-level as well as the group-level.

A violation of the Markov condition occurs when $p(X_1|Y_1) \neq p(X_1|Y_2)$ and $p(Y_1|X_1) \neq p(Y_1|X_2)$. We tested both pairs of judgments using Bayesian paired samples t-tests (see bottom of Table 7). We found 'extreme' evidence for a violation of the Markov condition in both comparisons. In particular, participants judged conditionals with matching causes (e.g., $X_1|Y_1$) to be greater than those with mismatching causes (e.g., $X_1|Y_2$). This result is the opposite of what is predicted by an "illusory correlation" where individuals "see" a correlation between two less-frequent features (Kutzner, Vogel, Freytag, & Fiedler, 2011; Eder, Fiedler, & Hamm-Eder, 2011; Fiedler, 2000). In our experiments features X_2 and Y_1 are the features with low base-rate probabilities (i.e., these features are described as occurring in "few" animals). If we had an illusory correlation in our data, then we would have $p(Y_1|X_1) < p(Y_1|X_2)$. However, this is not the case, rather we observe that $p(Y_1|X_1) > p(Y_1|X_2)$.

The bottom left panel in Figure 2 shows a scatter plot with the two different pairs of judgments testing Markov violations. As can be seen in the figure, most of the data falls below the 45 degree line of identity, suggesting Markov violations occur at the individual-level as well as the group-level. In particular, the data for the comparison of $Y_1|X_1$ and $Y_1|X_2$ (blue squares in the figure) fall below the identity line showing that the first judgment ($Y_1|X_1$) is judged to be greater than the second ($Y_1|X_2$). This is the opposite of what would be expected if people were demonstrating an "illusory correlation".

Anti-discounting behavior occurs when individuals fail to discount additional causes when one or more causes are known. For example, judging $p(Y_1|E_1, X_1)$ to be greater than or equal to $p(Y_1|E_1)$ implies a failure to incorporate known information about X_1 and discount Y_1. We tested anti-discounting using the two pairs of judgments listed in the bottom two rows of Table 7. We found 'anecdotal' to 'moderate' evidence in support of the null hypothesis, suggesting the events in each pair are judged to be the same. This is indicative of a failure to discount. The bottom right panel in Figure 2 shows a scatter plot with the two different pairs of judgments for anti-discounting. As can be seen in the figure, most of the data falls around the 45 degree line of identity, suggesting anti-discounting occurs at the individual-level as well as the group-level.

Next, we were interested in whether the effects discussed above are correlated. If different people adopt different mental representations in our task (corresponding to different models in our framework), then we might expect some of the effects to be correlated. In particular, our modeling framework makes the strong prediction that reciprocity and memorylessness should co-occur. These two effects arise from the same underlying property of the 2D model, namely the assumption that events are represented by one dimensional subspaces. In order to examine such relationships, we defined four different measures: ReciprocityScore, MemorylessnessScore, MarkovScore, and AntidiscountingScore. All of the measures are calculated on an individual-level. The first three scores are calculated by taking the average of the absolute differences of the pairs in Table 7 for a particular effect. For example, the ReciprocityScore is given by,

$$\frac{|(X_1|E_1 - E_1|X_1)| + |(Y_1|E_1 - E_1|Y_1)| + |(X_1|Y_1 - Y_1|X_1)|}{3} \tag{28}$$

where a larger score indicates a larger “violation” of reciprocity (i.e., evidence against the 2D model). For the MemorylessnessScore, we used all four pairs shown in the middle of Table 7. Similar to the ReciprocityScore, a larger score indicates a larger deviation from memorylessness (again evidence against the 2D model). For the MarkovScore, we used the pairs $\{X_1|Y_1, X_1|Y_2\}$ and $\{Y_1|X_1, Y_1|X_2\}$. A larger value on this measure indicates a larger violation of the causal Markov condition. The AntidiscountingScore uses the pairs in the bottom two rows of Table 7 and is defined as

$$\frac{(X_1|E_1 - X_1|E_1,Y_1) + (Y_1|E_1 - Y_1|E_1,X_1)}{2} \tag{29}$$

Note that for this measure, we did not take the absolute value of the differences because positive and negative differences indicate different results. A positive difference corresponds to discounting, which is normatively correct in this situation. A zero or negative difference is suggestive of anti-discounting. Because participants often use rating scales in different ways (e.g., some participants will use the full rating scale while others are more cautious), we also divided each participant's scores by the standard deviation of that individual participant's responses to take into account differences in how different participants employ the response scale range.

The correlations between the four effects are given in Table 8. As anticipated, there is a strong positive correlation between reciprocity and memorylessness, suggesting that individuals that display reciprocity also display memorylessness. This novel finding is in agreement with our modeling framework in which memorylessness and reciprocity co-occur in individuals using a 2D representation. There are moderately negative correlations between reciprocity and Markov violations as well as memorylessness and Markov violations. This implies that individuals showing more evidence for reciprocity and memorylessness (i.e., participants with smaller ReciprocityScores and MemorylessnessScores) also show more evidence for violations of the Markov condition. We also see a moderately negative correlation between memorylessness and anti-discounting. At first, this might seem surprising since memorylessness and anti-discounting are both non-normative and we might anticipate that they should be positively correlated. However, if an individual demonstrates memorylessness as in the 2D model, then they will judge $X_1|E_1,Y_1$ as $X_1|Y_1$ and $Y_1|E_1,X_1$ as $Y_1|X_1$ (that is, they will have no memory of E_1). Thus, the anti-discounting comparison becomes $X_1|E_1 - X_1|Y_1$ and $Y_1|E_1 - Y_1|X_1$. These differences will be positive whenever participants judge the relationship between the causes and effect to be stronger than the relationship between the two causes. Thus, it is quite reasonable to see a negative correlation between memorylessness and anti-discounting because greater memorylessness leads to the "forgetting" of the effect E_1 in the anti-discounting comparisons.

Modeling Results

The behavioral results of Experiment 2 provide at best ‘moderate’ evidence for reciprocity, memorylessness, and anti-discounting. However, there is ‘extreme’ evidence for violations of the Markov condition. Interestingly, we find a strong positive correlation between reciprocity and memorylessness, in line with predictions from our modeling framework. Overall, the results of Experiment 2 suggest that there are at least some participants that deviate from classical probability theory and can perhaps be modeled using representations containing incompatible events. In this section, we explore these individual differences further by comparing model fits to individual participant data. This is made possible because each participant provided a response based on a continuously varying rating rather than a choice between just three alternatives, so that the data and model can be compared directly. In particular we do not need to use the softmax function used for Experiment 1 to transform between probability and choice. Thus, the τ parameter is not used and each model has one less parameter for the fitting as compared to Experiment 1.

The judgments solicited in Experiment 2 were mainly concerned with testing the four effects of reciprocity, memorylessness, Markov violations, and anti-discounting. Because of this they do not explore the full space of predictions, so that the model predictions are independent of some of the parameters. This is mainly a problem for the 2D POVM model. One of the angles (we chose θ_E) can therefore be fixed in this model when applied to this data set. The remaining priors, and other details of the modeling, were as for Experiment 1. Note that there were a total of 14 questions per block, but only 12 of them were unique. We fit the 12 unique questions.

There are several ways to understand the results of the modeling. One is to simply ask for what proportion of participants was each model favored over the other two, according to the DIC? The results are that the 8D model was preferred for 40% (23/58) of participants, the $4D_{CC}$ model was preferred for 34% (20/58) of participants, and the 2D POVM model was preferred for the remaining 26% (15/58) of participants. The 8D model is therefore clearly the strongest single model, but equally more participants are better fit by a non-classical representation of some form than a classical one.

Assuming that these subgroups identified as being fit better by the different models are meaningful, this should be reflected in the behavioral analysis. Figure 3 shows the behavioral analysis of each of the four effects for each subgroup. The results are largely as expected, suggesting that the modeling has appropriately identified subgroups of participants displaying distinct behaviors. The 2D POVM group displays a lower ReciprocityScore (BF_{10} = 26.16; F(2,55) = 7.48, p = 0.001) as well as a lower MemorylessnessScore (BF_{10} = 13.09; F(2,55) = 6.57, p = 0.003). The 2D POVM group also appears to display a higher MarkovScore and higher AntidiscountingScore, but these results are not significant.

Another way to understand the modeling is to look at the fit between observed and predicted probabilities across all participants. This is shown in Figure 4. We can see that 8D and $4\mathrm{D}_{\mathrm{CC}}$ models perhaps provide tighter fits, but they also seem to systematically overestimate low probabilities and underestimate high ones. The probabilities have been split up into three types for each plot; probabilities of the form $p(X|Y)$ are shown by (blue) circles, probabilities of the form $p(E|X)$ or $p(X|E)$ are shown as (red) squares, and probabilities of the form $p(E|X,Y)$ or $p(X|E,Y)$ are shown by (green) triangles. A visual inspection of the plots seems to suggest that the different models over and under estimate different types of probabilities, and this does indeed appear to be the case. Bayesian paired samples t-tests reveal that the 8D model systematically underestimates probabilities of the form $p(E|X)$ or $p(X|E)$ by about 6.8% ($\mathrm{BF}_{10} \sim 10^{10}$). The $4\mathrm{D}_{\mathrm{CC}}$ model also underestimates these probabilities, by about 5.1% ($\mathrm{BF}_{10} \sim 10^{6}$) and overestimates probabilities of the form $p(E|X,Y)$ or $p(X|E,Y)$ by about 4.3% ($\mathrm{BF}_{10} \sim 8,000$). The 2D POVM model also underestimates probabilities of the form $p(E|X)$ or $p(X|E)$, but by a rather smaller margin of 3.6% ($\mathrm{BF}_{10} = 75$).

Overall the 8D and $4\mathrm{D}_{\mathrm{CC}}$ models do fit well for some participants, but they tend to systematically under/over estimate certain types of probabilities. In contrast the 2D POVM model is favored by the DIC for fewer participants, but it does have the advantage of being less systematically biased when evaluated over all participants.

Conclusions

In summary, it is reasonably clear that there are large individual differences in the best fit models to Experiment 2. Although some participants are reasoning in agreement with a classical representation, most (60%) are not. The fact that those using a quantum representation are split between two types ($4D_{CC}$ and 2D POVM) helps account for the fact that we saw no obvious preferred model emerge from the results of Experiment 1. However by fitting individual data we are able to identify three distinct groups that appeared to be better fit by the 2D POVM, $4D_{CC}$ and 8D models, with the behavioral data for these subgroups matching the qualitative predictions of each of these models.

It is less clear why there should be these individual differences in the best fit models. What are the factors that drive individuals to construct differing representations, and are these fixed, or can an individual adapt their representation over time?

Experiment 3

The results of Experiment 2 indicate that there are individual differences in the way that people judge causal events. This experiment aims to address the sources of those differences. In particular, our goal is to show that differences in performance in inference tasks can be related both to familiarity with the task and to individual differences in cognitive ability (i.e., how reflective participants are in the task). We hypothesize that simpler representations (such as those with incompatible events) may be adopted when reasoning in an intuitive, heuristic way, while more complex representations are only formed when reasoning in a more deliberative way. Thus, the type of representation used by an individual may be linked to a simple measure of cognitive style, the Cognitive Reflection Test (CRT), which measures an individual's ability to suppress an initial "gut" response that is incorrect, in favor of a deliberative correct response (Frederick, 2005). We also hypothesize that familiarity with a scenario may allow individuals to construct more complex representations, with more compatible events. Thus, we expect that repeated exposure to a scenario could lead to a fully classical representation of events. The inclusion of the CRT and a multi-block testing procedure are the major differences between Experiment 3 and the previous ones.

Methods

60 undergraduate students from a US university participated in the experiment online at a time of their choosing for course credit. The instructions, stimuli, and procedures were similar to Experiment 1. Participants were randomly assigned to one of two novel animal categories (either African Lake Shrimp or Kehoe Ants). The features of the animals and the relationships between features were described very similarly to Experiment 1. One difference between the experiments was the inclusion of probabilistic information about the feature base-rates in the current experiment. For example, in the African Lake Shrimp category participants were told that “Most shrimp (90%) have a high amount of ACh whereas a few (10%) a low amount of ACh”. In both categories, 90% of animals had feature X_1, 10% had feature Y_1, and 50% had feature E_1. The learning stage was identical to Experiment 1.

After completing the learning stage, participants completed six blocks of trials where they were asked to make decisions about the value of different features. There were two block types (BX and BY) that were repeated three times in an alternating fashion (e.g., BX_1, BY_1, BX_2, BY_2, BX_3, BY_3). Participants were randomly assigned to start with either the BX_1 or BY_1 block. Each block contained nine questions (see Table 9) where participants were asked to select the value of a particular feature. Similar to Experiment 1, participants were given three response options: feature value 1, feature value 2, or equally likely to be feature value 1 or 2. For example in the African Lake Shrimp category, the question ”What type of body weight do you think this shrimp has?”, had the following response options: 1) a low body weight, 2) a high body weight, and 3) equally likely to be low or high.

Some questions asked participants to make a sequence of choices about a feature value (e.g.,

E) as they learned new information about the other features (e.g., X and Y). As in the previous experiments, the information about the value of the first feature remained on the computer screen when new information about the second feature was presented. Thus, participants had all feature information available on the screen during the final choice. In the BX_i (BY_i) block, information about feature X (Y) was always presented before information about feature Y (X) in sequences involving both features. This helped reduce the influence of memory on future decisions about reverse orderings. An important consequence of the blocking is that order effects can only be evaluated by comparing responses across blocks (e.g., comparing $E|X_1,Y_2$ with $E|Y_2,X_1$ requires comparing responses across BX_i and BY_i). We repeated each block pair three times in order to examine the influence of task familiarity on responses. Thus, in the analyses below, we bin the blocks into pairs (BX_1 and BY_1, BX_2 and BY_2, BX_3 and BY_3).

After finishing the six blocks, participants completed the CRT (Frederick, 2005). This test assesses an individual's ability to suppress a spontaneous and intuitive ("System 1") wrong answer in favor of a deliberative and reflective ("System 2") correct answer. The test consists of three items using a free-response format and is scored by counting the number of correct responses across the items. Performance on the CRT has been correlated with many behavioral measures including temporal discounting, mental heuristics, and risk preferences (Frederick, 2005; Toplak, West, & Stanovich, 2011).

Behavioral Results

All participants were included in the analyses. The average score on the 10 question multiple choice test was 9.6 indicating most participants correctly learned the feature values and causal relationships during the first part of the experiment. Similar to Experiment 1, we calculated choice scores following Rehder (2014) by assigning the following values to the three response options: feature value 1 = 1, feature value 2 = 0, and equally likely = 0.5. Note that there were no differences between choices in the two different animal categories ($\mathrm{BF}_{10} = 0.089$) and so responses were collapsed for the following analyses. Table 9 shows the mean choice scores collapsed across repeated blocks along with standard deviations for each question.

For the following analyses, we grouped the blocks into pairs (i.e., first: BX_1 and BY_1, middle: BX_2 and BY_2, and last: BX_3 and BY_3). We first assessed order effects, reciprocity, memorylessness, and violations of the Markov condition by applying Bayesian paired samples t-tests, similar to Experiments 1 and 2. The results of these tests for each block pair are shown in Table 10. Overall, we see 'extreme' evidence for order effects similar to Experiment 1. We also have 'extreme' evidence against reciprocity and memorylessness. Note that this experiment examines "cause-cause" reciprocity and "cause-cause" memorylessness. In Experiment 2, "cause-cause" reciprocity was examined by comparing $X_1|Y_1$ and $Y_1|X_1$ and there was 'moderately strong' evidence for the effect. Thus, the 'extreme' evidence against reciprocity in this experiment is surprising. However, there is a very important difference between Experiments 2 and 3. Experiment 3 provides precise probabilities for the base-rate of features (e.g., 90% of animals had feature X_1). Previous work using scenarios similar to the ones in the present experiments has shown that causal judgments can be influenced by the presence of precise base-rate information (Rehder, 2003b). This is one possible explanation for the differences between these experiments. The 'extreme' evidence against memorylessness is similar to the results found in Experiment 1 for comparisons involving causes with mismatching features. In the first block pair, we see evidence for violations of the Markov condition (ranging from 'anecdotal' to 'strong' evidence). This is similar to the results found in Experiment 2, however the evidence was stronger in that experiment. The difference in the strength of evidence between the two experiments might be the result of the precise base-rate probabilities in Experiment 3 or the different response scales used in the experiments.

Next, we calculated the following four measures for each block pair: OrderScore, ReciprocityScore, MemorylessnessScore, and MarkovScore. Similar to Experiment 2 these scores are calculated by taking the average of the absolute differences of the pairs in Table 10 for a particular effect. For example, the OrderScore is given by,

$$\frac{|(E|X_1,Y_2)-(E|Y_2,X_1)|+|(E|X_2,Y_1)-(E|Y_1,X_2)|}{2} \tag{30}$$

where larger scores indicate larger order effects. We also grouped individuals into three groups based on CRT scores: high = CRT score of 3, medium = CRT score of 1 or 2, and low = CRT score of 0. We combined individuals with CRT scores of 1 and 2 into a single group so we had roughly an equal number of individuals per group. There were 21 participants in the CRT high group, 19 in the CRT medium group, and 20 in the CRT low group. Using the four measures for each of the three CRT groups, we can examine differences in the effects due to differences in reasoning ability as well as how the effects change with experience gained through exposure.

The top left panel of Figure 5 shows OrderScore for the three CRT groups across the block pairs. A Bayesian repeated measures ANOVA showed that a model including both block pair and CRT group (but no interaction term) was preferred to all other models ($\text{BF}_\text{M} = 7.52$)[4] as well as the null model ($\text{BF}_{10} = 464.67$). These results were corroborated by a traditional repeated measures ANOVA that showed a main effect of block pair ($F(2,118) = 9.86$, $p < .001$) and CRT group ($F(2, 57) = 3.45$, $p = 0.039$). The interaction was not significant. From Figure 5, we see that the low and medium CRT groups have a larger OrderScore than the high CRT group and that OrderScore decreases across block pairs. This implies that order effects are larger for low and medium CRT groups, but decrease with experience.

The top right panel of Figure 5 shows ReciprocityScore for the three CRT groups across the block pairs. A Bayesian repeated measures ANOVA showed that a model including the interaction of block pair and CRT group was preferred to the null model ($\text{BF}_{10} = 3.04$). These results were corroborated by a traditional repeated measures ANOVA that showed a significant interaction between block pair and CRT group ($F(4, 118) = 3.22$, $p = 0.015$). The main effects were not significant. From Figure 5, we see that in the first two block pairs, the low CRT group has a lower

ReciprocityScore than the medium and high CRT groups. Further, ReciprocityScore for the low CRT group increases across block pairs showing increasing deviations from reciprocity with experience.

The bottom left panel of Figure 5 shows MemorylessnessScore for the three CRT groups across the block pairs. A Bayesian repeated measures ANOVA showed that a model including CRT group was preferred to all other models ($BF_M = 5.70$) and was favored over the null model ($BF_{10} = 2.13$) These results were corroborated by a traditional repeated measures ANOVA that showed a marginally significant main effect for CRT group ($F(2, 57) = 3.13$, $p = 0.052$). There was no main effect of block pair and the interaction was not significant. From Figure 5, we see that the low and medium CRT groups have a smaller MemorylessnessScore than the high CRT group.

The bottom right panel of Figure 5 shows MarkovScore for the three CRT groups across the block pairs. A Bayesian repeated measures ANOVA showed that a model including CRT group was preferred to all other models ($BF_M = 5.49$) and was favored over the null model ($BF_{10} = 2.00$) These results were corroborated by a traditional repeated measures ANOVA that showed a significant main effect for CRT group ($F(2, 57) = 3.47$, $p = 0.038$) and a slight interaction between block pair and CRT group ($F(4, 118) = 2.10$, $p = 0.086$). There was no main effect of block pair. From Figure 5, we see that in the first two block pairs, the low CRT group has a higher MarkovScore than the medium and high CRT groups. Further, MarkovScore for the low CRT group decreases across block pairs showing a reduction in violations of the Markov condition with experience.

Modeling Results

For the modeling we used the CRT score and block pairs to split the data into 3 (CRT groups) $\times$ 3 (block pairs) sets, from Low CRT in the first blocks to High CRT in the last blocks. This allows us to examine model performance as a function of CRT group and also see how the various models perform for each group as participants proceed through the task. Each block pair had a total of 18 questions (9 per block). Because some questions were repeated, there were a total of 15 unique questions. We fit these 15 questions for each CRT group and block pair. All priors, and other details of the modeling, were as for Experiment 1.

The results are shown in Figure 6. A number of features stand out; firstly the 2D POVM (solid line) initially performs much better than the other models for the low and medium CRT groups. In contrast the $4D_{CC}$ (dashed line) and 8D (dotted line) models initially perform better in the high CRT group. This is strong evidence that the mental representation participants construct when first presented with the scenarios depends on their CRT measure, and therefore on whether they are primarily engaging in spontaneous and intuitive ("System 1") thinking or deliberative and reflective ("System 2") thinking during the task. The more 'classical' $4D_{CC}$ and 8D models seem to be associated with more deliberative and reflective thinking, which fits well with the fact that quantum models are often used to explain decision making which is inconsistent with (classically) normative prescription and appears to be driven by heuristics (Busemeyer et al., 2011).

The other obvious feature of this data is that the ability of the various models to capture the data from the low CRT group varies dramatically as the experiment progresses. In the initial blocks the 2D POVM model vastly outperforms the other two models, whereas by the final two blocks all three models are performing at about the same level. This implies a significant shift in representation through the course of the experiment for this low CRT group, such that the representation these participants are using to reason about the scenarios becomes much more classical with repeated exposure. In the General Discussion section, we outline some possible explanations for this.

In contrast, the performance of the three models does not seem to vary that much across blocks in the medium and high CRT groups. In the high CRT group this is less surprising, since one would expect additional experience to continue to favor the 8D or the $4D_{CC}$ models, but is unclear why the 2D POVM model remains the preferred one even towards the end of the experiment for the medium CRT group. However, note that in the low and medium CRT groups, the $4D_{CC}$ and 8D models in the last block pair have similar DIC values. Specifically, the DIC for the $4D_{CC}$ model is -12.35 for the medium CRT group as compared to -14.23 for the low CRT group in the last blocks. Likewise, the DIC for the 8D model is -9.65 for the medium CRT group as compared to -13.07 for the low CRT group in the last blocks. Thus, there is not much difference in the $4D_{CC}$ and 8D model fits for the medium and low CRT groups in the last blocks. The main difference between these two groups is in the fit of the 2D POVM. In addition it is somewhat surprising that no clear winner emerges between the 8D and $4D_{CC}$ models, even towards the end of the task. Some potential reasons for this are outlined in the General Discussion.

It is worth noting that for the 2D POVM model in the low CRT condition the value of the parameter in the POVM, ε, starts at a relatively low level of 1.0% for the first blocks, before increasing to 2.6% by the last blocks (in contrast the value for the medium CRT group is stable at around 3.0% across all blocks.) Recalling that the 2D POVM model reduces to the simpler 2D model for small enough ε, we can see that initially the Low CRT group are best fit by a model very close to the simplest possible, 2D, quantum model, suggesting they possess a particularly simple mental representation. However, with an increase in exposure to the task, or a higher CRT level, the best fitting 2D POVM model moves away from this.

In Figure 7 we give examples of three of the fits, one from the low CRT group in the initial stages of the experiment, where the 2D POVM model fits best, and two from the High CRT group in the later stages of the experiment, where the $4D_{CC}$ and 8D models perform about as well as each other. Overall each model gives a good fit to the relevant data, although it is notable that there are still small order effects visible in the data for the high CRT group in the later blocks of the experiment, as noted in the behavioral results above.

Conclusions

The behavioral results of Experiment 3 show that participants' choices appear to change as they gain familiarity with the task. We also found evidence that individuals displaying non-classical effects tend to score lower on the CRT. In particular, the low CRT group tends to show larger order effects, larger violations of the Markov condition, stronger reciprocity, and memorylessness. According to the CRT, these participants are likely using an intuitive ("System 1") style of thinking when completing the task. Interestingly, these individuals seem to improve the most with repeated exposure to the scenarios.

The modeling demonstrates that performance on the CRT task is a good predictor of which model in the hierarchy will provide the best fit to the experimental data at the beginning of the task. It also shows that for the low CRT group model fits change as a function of familiarity, with the 2D POVM model providing a worse fit over time. However there are still some intriguing results which are harder to explain, particularly why neither experience with the task nor CRT grouping seem to distinguish between the $4D_{CC}$ and 8D models. We return to this question in the next section.

General Discussion

This paper presents an alternative approach to constructing models of human inference. Rather than elaborating an existing rational model (for example, adding nodes and edges to a CGM), we expand the possible representations used to perform inference. We propose using quantum probability theory to construct a hierarchy of mental representations. These models are constructed using different configurations of compatible and incompatible events. Compatible events are identical to the ones in classical probability theory. Incompatible events are unique to quantum theory. If two events are incompatible, their joint event does not exist and they must be processed sequentially. Crucially, within a quantum approach, there is a principled framework for integrating representations of varying degrees of compatibility/ incompatibility. Different levels in the hierarchy correspond to models of different dimensionality. Low dimensional models represent lower levels in the hierarchy and correspond to simple mental representations (that is, representations with few joint events). Higher dimensional models represent higher levels in the hierarchy and use more joint events, providing a more complex representation of causal events.

Models at the highest level in the hierarchy are equivalent to classical probability models (such models allow all possible joint events).

Throughout our lives, we are faced with a large variety of inference problems. In some situations, we have extensive knowledge of events. In other cases, there is only vague information about events. Also, different individuals might adopt different approaches to solving problems, sometimes relying on intuitive or spontaneous reasoning over more deliberative approaches. Different models in our hierarchy might be used for different types of inference problems and by different individuals. In situations where individuals do not have a lot of past experience with a problem, it seems reasonable to expect that a simple mental representation with few joint events might be adopted. However, when individuals have extensive knowledge of events or information about a problem is very clear, individuals might form a more complex representation of events. Mental representations might also be tied to cognitive ability where more deliberative reasoning leads to the formation of more complex compatible representations.

Psychological implications

A critic might wonder what new psychological insights are gained by adopting the approach we propose in this paper. We feel that there are many new psychological questions that can be answered using the proposed quantum framework. First, as far as we are aware, our framework is the first to suggest that a previously disparate set of phenomena (order effects, reciprocity, memorylessness, violations of the Markov condition, and anti-discounting) potentially arise from the same underlying set of principles. Specifically, when an individual has a partial mental representation of a problem (as modeled using incompatible events), one can expect most or all of these effects. This opens up new avenues of inquiry within psychology because previous research has tended to explore non-normative phenomena in isolation. For example, order effects and reciprocity are rarely studied in causal inference (although order effects have been studied in the related domain of causal learning, see for example Dennis & Ahn, 2001; Collins & Shanks, 2002;

Abbott et al., 2011). However, if our hypothesis is correct that order effects and reciprocity are likely to co-occur with Markov violations and anti-discounting, then this suggests experiments in causal inference should also examine these phenomena. Without this prediction from the present modeling framework, it is unlikely researchers would ever think to look for these effects in causal inference.

More generally, this framework integrates normative and non-normative influences into a single unified theory. Previously, such influences have been considered as corresponding to entirely different mechanisms. In the present work, we show that such influences reflect the same kind of probabilistic inference, but on different (compatible versus incompatible) representations. We explain how non-normative influences can be understood using the single idea of incompatible representations. In our approach, different non-normative influences can emerge from differences in the compatibility of events.

The present approach also made new predictions regarding the influence of practice and individual differences, which go beyond existing frameworks. Even though the issue of individual differences has been considered before, we provide an interesting hypothesis (supported by the data) which uniquely emerges from the present formalism. This is achieved because the framework formalizes the idea of ‘simple mental representations’ as compared to ‘complex mental representations’. Without a formal definition of these terms it is impossible to answer questions such as “How does cognitive thinking style relate to mental representations?” or “Does familiarity with a problem change a person’s mental representation of the problem?” In order to answer these questions, one first needs to define what a mental representation is and how it might vary among individuals. That is exactly what the present modeling framework accomplishes. Further our framework provides a precise way in which different mental representations are related - through the inclusion or exclusion of joint events. This opens the door to the question of how people transition between representations. In the section below, we sketch out one possibility for how this might occur. With a better understanding of how transitions occur, we can develop new ways to train people in order to improve their judgments and decision-making.

Transitioning between representations

The results of Experiment 3 suggest the intriguing possibility that participants change their representation of the events as they gain familiarity with the task, gradually becoming less 'quantum' and more 'classical', in the specific sense of moving from a totally incompatible representation (2D POVM) either to a totally compatible one (8D), or at least to one where the events X and Y (the two causes) are compatible (4D_{CC}).

However this raises a number of questions. How are we to understand the process of transitioning from, e.g. a 2D to a 4D representation? What might participants 'learn' to cause them to make this transition? Why is it that for the high CRT group there does not appear to be any transition from the 4D_{CC} to the 8D representation, if the latter is supposed to be more classical? In this subsection we provide some preliminary analyses showing how we might express compatible and incompatible representations in the same language, and what has to change to affect a transition between the two, and we argue that the results of Experiment 3 are in line with these general expectations.

Different models in our hierarchy are embedded in different Hilbert spaces, which makes it difficult to think about transitioning between them. An alternative approach is to work in the language of quasi-probabilities (see e.g., Halliwell & Yearsley, 2013, and references therein). Quasi-probabilities function like standard probabilities, except that some elements may lie outside of the range $[0,1]$. In this sense quasi-probabilities formalize the idea that probability distributions 'do not exist' in some cases. For example, we might have a quasi-probability distribution $q(X,Y)$ where the marginals are genuine probability distributions, e.g. $\sum_Y q(X,Y) = p(X)$ etc, but some of the elements of $q(X,Y)$ are negative.

In our experiments participants provided probabilities such as $p(X)$ and $p(E,Y)$. An explanation for these judgments based on classical probability theory exists if and only if these probabilities can be expressed as the marginals of some joint probability distribution $p(X,Y,E)$. If such a joint probability distribution does not exist, it is still possible to define a quasi-distribution $q(X,Y,E)$ such that $p(X), p(E,Y)$, etc. are given by the appropriate marginals. In other words, a classical representation of the events X, Y, and E exists if and only if the quasi-probability distribution $q(X,Y,E)$ is in fact positive. Therefore we can think about transitioning from an incompatible to compatible representation as equivalent to transitioning from a quasi-distribution to a genuine probability distribution.

For brevity we focus on the case of two variables X and Y. Suppose they can take values $X = \pm 1, Y = \pm 1$. Then the joint probability may be written (Halliwell & Yearsley, 2013; Halliwell, 2014),

$$p(X = x_i, Y = y_j) = \frac{1}{4}\left(1 + x_i(2p(X = 1) - 1) + y_j(2p(Y = 1) - 1) + x_i y_j C_{xy}\right) \tag{31}$$

where

$$\begin{aligned} C_{xy} &= p(X = 1, Y = 1) + p(X = -1, Y = -1) - p(X = 1, Y = -1) - p(X = -1, Y = 1) \\ &= p(\text{'same'}) - p(\text{'different'}) \end{aligned} \tag{32}$$

is the correlation function.

It is easy to see that $p(X = x_i, Y = y_j)$ has the expected marginals, and that they do not depend on C_{xy}. Given $p(X)$ and $p(Y)$ therefore, different values for C_{xy} will lead to different joint distributions, some of which will be probability distributions, and some of which will only be quasi-distributions.

For any marginals $p(X)$ and $p(Y)$ it is always possible to choose a C_{xy} such that $p(X = x_i, Y = y_j)$ is a probability distribution, however this is not in itself very interesting. A more interesting approach is to imagine that we guess a value for C_{xy} given some information we have about the problem, and we ask, given $p(X), p(Y)$ and our guess for C_{xy}, does this lead to a compatible representation for X and Y?

It is helpful to work with the variable,

$$S_{xy} = \frac{1}{2}(1 + C_{xy}) \tag{33}$$

which ranges from 0 if the events are perfectly anti-correlated, to 1 if the events are perfectly correlated. We argue that it is reasonable therefore to equate, S_{xy} with the similarity $Sim(X,Y)$ between the two events. In psychological terms, this means when given $p(X)$ and $p(Y)$ and asked to make judgments about the joint events, participants 'fill in the blanks' using the similarity. It turns out that in the 2D case the expression for S_{xy} is exactly equal to the quantum expression for the similarity (Pothos, Busemeyer, & Trueblood, 2013; Pothos et al., 2015).

Now consider a simple example of two variables where $p(X = 1) = 0.9$ is high, but $p(Y = 1) = 0.1$ is low. Further suppose that $Sim(X,Y) = 0.7$ is judged to be high, perhaps because they are both causes of some other event. From Eq.31 we can see,

$$\begin{aligned} p(X = -1, Y = +1) &= \frac{1}{2}\left(1 - p(X = 1) + p(Y = 1) - Sim(X,Y)\right) \\ &= \frac{1}{2}(1 - 0.9 + 0.1 - 0.7) = -0.25 < 0 \end{aligned} \tag{34}$$

so that $p(X = x_i, Y = y_j)$ is not a valid probability distribution. It is easy to see what is going wrong here, X and Y cannot be highly correlated given the marginals, and so assuming this makes it impossible to form a sensible joint distribution. Indeed given $p(X = 1) = .9$, $p(Y = 1) = .1$ the joint probability exists only if $Sim(X,Y) \leq 0.2$. Even a moderate guess for the similarity will make it impossible to construct a joint distribution.

To recap, given only the marginals $p(X)$ and $p(Y)$, construction of a joint probability distribution is equivalent to fixing the correlation function C_{xy}. Participants with no prior knowledge might plausibly do this with the aid of the perceived similarity between X and Y, setting $C_{xy} \approx 2 \times Sim(X,Y) - 1$.

This suggests how we might view the process of changing representation. Suppose participants initially use a heuristic to set C_{xy} by relating it to similarity. Some of the elements of the quasi-probability distribution might therefore be negative, or equivalently a classical representation of X and Y is impossible. However experience with the task may serve to teach participants about the correct relationship between X and Y. Participants should therefore revise their estimate of C_{xy} away from the initial value, and eventually, C_{xy} will lie in a range where all the elements of the quasi-distribution $q(X,Y)$ become positive, and thus a classical representation of these beliefs about X and Y is possible.

We hypothesize that this is what happens in Experiment 3. Initially participants may assume X and Y are highly correlated, since the only information they have about these events is their similarity in producing the effect E. However every time we ask participants to judge $p(E|X,Y)$ we do so by telling them we have collected a sample organism which has either features X_1,Y_2 or X_2,Y_1. Therefore we are, unintentionally, teaching participants that X and Y are anti-correlated. Participants should be able to use this information to revise their estimate of the correlation between X and Y, and if they do this sufficiently a compatible representation may be possible.

If this explanation is correct we should see large changes in judgments of the conditional probabilities between X and Y over the course of Experiment 3 for the Low CRT group, and indeed this is what we find. The largest changes in the judged probabilities between the first and last block pairs are for $p(X_1|Y_2)$ and $p(Y_1|X_1)$. In addition, the largest difference between the initial probability judgments of the Low CRT group and the other CRT groups is for these same probabilities. This suggests that much of what makes the Low CRT group unique is their initial belief about the correlation between X and Y. This also explains why transitions between the $4\mathrm{D}_{\mathrm{CC}}$ and 8D models do not seem to occur - both have X and Y compatible, and so learning about the relationship between X and Y will not cause a transition between these two representations. These models differ in whether E,X or E,Y are compatible, but no information that might help participants refine their view of the relationships between the causes and the effect is presented.

We have outlined a framework that can account for transitions between incompatible and compatible representations in terms of learning specific types of information about the correlation between events. This framework is consistent with our hierarchy of representations, future work should explore this further, with a view to designing experiments where transitions between representations might be induced by presenting specific types of information.

Strategy sprawl

The modeling framework presented here is an example of a ‘cognitive toolbox’ model. Toolbox models are popular is many areas of psychology including linguistics (Eisenberg & Becker, 1982), decision-making (Gigerenzer & Todd, 1999), development (Coyle, Read, Gaultney, & Bjorklund, 1998), categorization (Busemeyer & Myung, 1992), and causal reasoning (Rehder, 2014), to name a few. The basic idea behind all of these approaches is that individuals are equipped with a set of strategies to solve a given task. For example, Rehder (2014) proposed a set of five models to explain how people reason about causal events (which we discuss in more detail below). While these modeling frameworks provide a powerful approach to understanding human cognition, they all face a similar problem: strategy sprawl. As the number of models or strategies in the toolbox grows, the framework becomes increasingly flexible.

While the hierarchy of models presented here is not immune to the issue of strategy sprawl, there are important constraints on how the framework can grow. First, the models in our framework must all obey the axioms of quantum (or classical) probability theory. This provides a strict set of rules underlying all of the models. It is not the case that we can pose any arbitrary model and add it to the framework. Given a specific problem domain, such as the common effect network discussed in the present paper, the number of possible models is limited. There are only a handful of ways to construct different probabilistic models for this particular problem.

In addition, recent advances in comparing toolbox models using hierarchal Bayesian methods show great promise (Scheibehenne, Rieskamp, & Wagenmakers, 2013). In this approach, different models (for example, toolbox models with different numbers of strategies) are compared using Bayes factors, which naturally account for the trade-off between model complexity and explanatory power. We believe this approach could be applied to our hierarchy of models and we see this as an excellent direction for future research.

Alternative models

Alternative modeling approaches can also explain some of our experimental results. In particular, Rehder (2014) conducted a large investigation of causal inference, examining both CGMs and non-normative reasoning strategies. He focused on three non-normative models: a Conjunctive Model (CONJ), a Shared Disabler Model (DISAB), and an Associative Model (ASSC). The CONJ model assumes that conditional probabilities are evaluated conjunctively. That is, conditional probabilities such as the probability of an effect (e.g., high body weight) given a cause (e.g., an accelerated sleep cycle) are instead evaluated as joint probabilities (e.g., high body weight AND accelerated sleep cycle). The second strategy, DISAB, assumes a hidden disabling mechanism by introducing an additional variable imagined by participants. This additional variable probabilistically influences one or more of the existing causal relationships. For example, a participant might imagine an additional variable, serotonin levels, that might probabilistically moderate the causal relationship between sleep cycle and body weight. The third strategy, ASSC, assumes an associative Markov random field. Essentially, this assumes a correlational relationship between variables, without allowing for any specific direction of causality.

Rehder (2014) found that no single model could adequately account for his data, a conclusion which closely resonates with ours. As with Rehder (2014), the question then becomes what is the appropriate mixture of strategies or models to understand the relevant human behavior. Rehder (2014) reported that a linear combination of all four strategies (CGM plus the three non-normative models) provided a good account of his data. Thus, he concluded that causal reasoning must involve both normative and non-normative influences. Rehder explored this further by identifying two groups of participants, called ‘associative’ and ‘causal’ reasoners, the former group displaying insensitivity to causal direction and the latter group displaying more normative behavior. Overall, Rehder’s approach is promising as it can account for individual differences in some effects (such as violations of the Markov condition). However, it is less clear how his modeling approach could cover other phenomena such as order effects, reciprocity, and memorylessness.

As noted, the normative component in Rehder’s (2014) approach, based on the CGM, is a special case of the 8D fully Bayesian model in the proposed hierarchy. However, further equivalences are hard to establish. The CONJ and ASSOC models use ad hoc probabilistic rules (neither classical or quantum) and the DISAB model adds extra variables in the CGM. Despite the fact that using Rehder’s (2014) four models together can lead to good descriptive results at least for his data, we think there are two important merits of the present approach. First, our approach covers more phenomena than discussed in Rehder (2014). In particular, our approach can explain order effects, reciprocity, and memorylessness, which are rarely studied in casual inference, and why these effects co-occur with Markov violations and anti-discounting. Second, our framework provides a detailed description of how the different components in the hierarchy are linked, because they are all based on essentially the same overarching probabilistic framework (of quantum theory, which is reduced to Bayesian theory when employing a fully compatible representation). In our hierarchy of models, moving from the lowest level to the highest involves changing assumptions about the compatibility of events. As one moves up the hierarchy, more joint events are included. Rehder’s approach lacks a theory about how the different strategies are connected. Thus, it seems doubtful that Rehder’s approach could account for the practice effects we find in Experiment 3. Our framework provides a specific prediction about how representations are related to familiarity, which could eventually lead to methods for improving people’s judgments through training.

Notwithstanding the above points, it is clearly a computational issue whether Rehder’s models can predict behavior which cannot be accommodated by the present approach. The study of this important issue cannot be resolved easily, because it requires examining participant behavior at the level of individual differences. The extensive corresponding analyses (employing a hierarchical Bayesian individual differences approach) are reported in Mistry, Trueblood, Vandekerckhove, and Pothos (submitted). Using the combined data from four experiments presented in Rehder (2014) (a total of 315 participants), Mistry et al. (submitted) compared a fully incompatible quantum model to the five models discussed in original paper: a CGM, the three non-normative models CONJ, DISAB, and ASSC, and the weighted average of the individual models. Hierarchical Bayesian model comparisons showed that the quantum model was preferred to all five Rehder models for three different causal network structures (chain, common cause, and common effect). Further, the quantum model predictions had the highest correlation with the observed data (between 0.87 and 0.93 for the three network structures). We view these results as strong evidence that the quantum approach can generalize to a wide range of human behavior.

Relationship to 'structurally local' causal inference

Our hierarchy of models can be viewed as a generalization of the idea that inference is achieved through *local computations* (Fernbach & Sloman, 2009). This theory argues that causal inference is *structurally local*. That is, when people are faced with a complex problem, they break the problem up by focusing on pairs of events rather than all events simultaneously. Inferences about the complete problem are constructed by combining local inferences piece by piece. The original version of the structurally local hypothesis most closely relates to the $4D_{IC}$ model where individual cause-effect relations are compatible, but separate relations are incompatible. Inference about compatible events are equivalent to those from classical probability theory. However, inferences about incompatible events must be considered piece by piece. Incompatible events form separate sample spaces that are "pasted together" by unitary transformations and inferences are obtained by the serial evaluation of events.

Our approach generalizes the structurally local hypothesis by suggesting different ways people might break up a complex inference problem. In the 2D model, individuals break up the problem into the smallest possible pieces. That is, they can only focus on one variable at a time and inferences about the full structure are performed by piecing together these individual inferences. Conceptually, this model shares ideas with hypothetical thinking theory (Evans, Handley, Neilens, & Over, 2007; Elqayam & Evans, 2013), which is a soft Bayesian approach to complex reasoning tasks. A central principle of the theory is that people focus on only one hypothesis at a time in their hypothetical thinking. The 2D POVM model is similar to the 2D model except it relaxes some of the assumptions in the 2D model and so breaks the strict properties of reciprocity and memorylessness. The 4D models capture the idea that individuals break up problems into several smaller (classical) "chunks" each involving two variables. In the 8D model, there are no local computations because individuals form a complete representation of the problem. This highest level of the hierarchy is classical.

Explanatory scope of the quantum framework

Understanding the explanatory level intended for a psychological model is important as it partly informs which comparisons with alternative models are meaningful. The predominant framework regarding levels of explanation is still that of Marr (1982), though the ideas from Griffiths et al. (2010) have been increasingly influential too. To understand the placement of the quantum framework for probabilistic inference, we briefly review the corresponding literature. Marr (1982) proposed three levels of explanation. Those presently relevant are the computational level and the algorithmic level, which is intended as intermediate between the computational level and the implementation level (the latter is about neurological/ biological processes and is not relevant here). The computational level concerns the *what* and the *why* questions for the system that is studied, that is "what is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out" (Marr, 1982, p. 25). Following Marr's example of studying a cash register, the what question is that the machine does arithmetic and the why is 'why arithmetic' instead of, for example, multiplication. A computational level specification of the cash register machine can then be elaborated easily, for example, given the objectives of the machine (which is to provide accurate measurement of the sum cost of a set of goods).

By contrast the algorithmic level concerns a process explanation of the studied system, specifically the representations that are employed by the system and the algorithms that operate on the representations to produce an output from an input. In other words, the algorithmic level is about the *how* question regarding the studied system. In the cash register example, an algorithmic level model could involve Arabic numerals for representation and the usual arithmetic calculus to carry out the necessary mapping from input to output information. Of course, the choices for algorithm and representation constrain each other (Anderson, 1978; Marr, 1982), and even then there may be alternative algorithmic level explanations which work.

A somewhat subtle point is that normative cognitive theories have to be computational level theories (Medin & Bazerman, 1999), but there are computational level theories that are not normative. Indeed, Marr (1982) does not reference normative prescription in his analysis. The first part of the assertion is straightforward: a normative decision theory would purport to provide a framework for why decision makers should be reasoning in a certain way - and so its explanatory objective is naturally at the computational level. However, there are computational level theories that are not normative. For example, consider the representativeness heuristic (Tversky & Kahneman, 1983), the idea that observers evaluate probability judgments through a similarity measure. The representativeness heuristic is a computational level explanation, since the emphasis is on the goal of the computation (compute probabilistic likelihood in terms of similarity) and on the appropriateness of the similarity heuristic (taking an extreme view for illustration, one could say that probability calculus is irrelevant and similarity supports probabilistic judgments as a readily available alternative cognitive mechanism). So, even though in decision-making, it seems natural to tie up the 'why' of a computational level model with a normative argument, this does not have to be the case.

Marr’s (1982) analysis has been hugely influential in psychology, but there are indications that it may be less suitable for modern cognitive models. The problem is that with an increasing sophistication of cognitive models, specification at multiple (Marr) levels is becoming more and more of a requirement. For example, consider the Leaky Competing Accumulator (LCA) model (Usher & McClelland, 2001), which concerns the time course of a perceptual choice task. It seems clear that the main objective of the model is at the algorithmic level (the specific way in which evidence accumulation at different time steps eventually leads to a decision). However, there are several overarching principles of the LCA model (e.g., the idea that evidence accumulation is leaky and a principle of lateral inhibition) that belong to the computational level.

These interpretational problems relate to probabilistic cognitive models and indeed Griffiths et al. (2010, p. 357) provided a well-thought generalization to Marr’s (1982) analysis by noting that “probabilistic models of cognition pursue a top-down or ‘function-first’ strategy, beginning with abstract principles that allow agents to solve problems posed by the world - the functions that minds perform - and then attempting to reduce these principles to psychological and neural processes.” Griffiths et al. (2010) further explain the intention of a probabilistic cognitive model with a range of questions, such as in relation to the information needed, the necessary representations, and the constraints on the computation. While most of these questions are computational level explanations, it is also clear that some of them are algorithmic level ones. Part of the problem is this: a probabilistic cognitive model (classical or quantum) embodies a default algorithmic assumption, namely the native probabilistic calculus. However, there is rarely a strong corresponding commitment. Indeed, when Sanborn, Griffiths, and Navarro (2010) considered the practicalities of computing Bayesian probabilities (e.g., priors), they suggested that the appropriate algorithmic level description of a Bayesian categorization model should involve various approximations regarding probabilistic computation.

Note also that the consideration of classical probabilistic levels often goes hand in hand with claims of optimality in cognitive process (Griffiths, Chater, Norris, & Pouget, 2012), thus apparently conflating a characterization of a model as normative and at the computational level of explanation. But, this is just because classical probabilistic models have been advocated in this particular way, given that normative justifications emerge particularly naturally for such models (e.g., see Oaksford & Chater, 2009); as argued above, even in decision making one can identify non-normative computational level models.

The quantum approach to probabilistic inference is a probability framework for human inference judgments, analogous to classical/ Bayesian models of inference. So, its explanatory objective is best stated using Griffiths et al.'s (2010) approach, as a model focusing on the mathematical principles that characterize human inference, when participants do not adhere to classical probability constraints. However, the quantum model can neither be assumed to be optimal nor normative. In all the scenarios we are employing, the representations that are at stake are (objectively) compatible (e.g., they concern biological properties of organisms). Therefore, if participants are representing these (objectively) compatible events in an incompatible way, then at best their reasoning process can be thought of as an instance of bounded rationality (Simon, 1955), on the further assumption that quantum representations represent a heuristic or suboptimal approximation to the classical ones. This latter assumption could be justified, because classical representations arguably impose high demands on attention and memory, since a high dimensional representation space needs to be constructed for all available variables (the requirement that a complete joint probability always exists means that it has to be possible to evaluate all variables concurrently). For participants not sufficiently familiar with the variables or not able or willing to sufficiently concentrate on the task (as might be evidenced by low CRT scores), an impoverished representation is created, which approximates the normative one to different degrees, but will also misrepresent some questions/ variables as incompatible. Clearly, this is not a complete picture, since it presupposes a specific idea for how (sufficient) attention and memory resources can lead to fully compatible representations, nevertheless it seems a reasonable speculation for how process limitations can make quantum representations relevant in human inference.

Overall, from the perspective of Griffiths et al.'s (2010) classification, the explanatory status of the quantum model is not unlike that of the heuristics proposed by Rehder (2014), though note that the explanatory emphasis of these heuristics does vary a bit between computational principles (computational level) and process (algorithmic level). The advantage of the quantum model is that there is a single set of coherent principles that capture the range of heuristic behaviors. Note, recent work indicates that quantum principles can be justified in a normative way too (Pothos, Busemeyer, Shiffrin, & Yearsley, in press). However, the conditions for when this is the case are not straightforward and, in any case, do not apply in the present causal inference experiments.

Finally, if one adopts the perspective of Marr's (1982) explanation levels, the quantum framework is about what is computed (briefly, a choice preference/ perception of likelihood) and why the computation has the form it does (briefly, information from several variables needs to be combined; in some cases this leads to incompatible representations, which imply quantum probabilities). The quantum approach does involve some algorithmic level assumptions, which are the probabilistic calculus native to quantum probability theory. However, such assumptions are not essential to the model, e.g., it is possible that quantum probabilities are computed via a diffusion-style process (broadly constrained by the relevant probability rules). A more pertinent question is how the quantum framework can be extended with algorithmic level assumptions, concerning for example memory or attention processes for how the relevant information is processed, in a way that enables a clearer picture of when to expect quantum and classical representations. Specifically, as noted, it is likely that memory or attention limitations may lead to quantum representations, when the classical ones are normative. In future work we will address this issue. The present objective is to firmly establish the relevance of quantum principles as a (computational) level description of human inference and how quantum principles can be related to the Bayesian (normative) principles for inference within the same broad probabilistic framework (Griffiths et al., 2010).

Concluding comments

Overall, our work sheds light on why classical models have been successful in many situations (e.g., CGMs), but can sometimes fail to agree with behavior in other situations. Equally, it helps us understand why quantum probability models can sometimes be successful, but are superfluous in other cases. Inference is neither inherently classical or quantum, but rather is tied to the representation of events constructed by the reasoner. For novel scenarios or when thinking intuitively, representations of events may be incompatible and quantum models are appropriate, however experience or more deliberative reasoning can lead to the formation of more complex compatible representations, which support classical models.

References

Abbott, J. T., Griffiths, T. L., et al. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2950–2955).

Aerts, D., Gabora, L., & Sozzo, S. (2013). Concepts and their dynamics: A quantum-theoretic modeling of human thought. *Topics in Cognitive Science*, *5*(4), 737–772.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249-277.

Busch, P., Grabowski, M., & Lahti, P. J. (1995). *Operational quantum physics*. Springer.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, *121*(2), 177-194.

Busemeyer, J. R., Pothos, E., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*, 193-218.

Busemeyer, J. R., Wang, Z., Pothos, E. M., & Trueblood, J. S. (2015). The conjunction fallacy, confirmation, and quantum theory: comment on tentori, crupi, and russo (2013). *Journal of Experimental Psychology: General*, *144*(1), 236-243.

Busemeyer, J. R., Wang, Z., & Trueblood, J. S. (2012). Hierarchical bayesian estimation of quantum decision model parameters. In J. R. Busemeyer (Ed.), *Qi 2012, lncs 7620.* Berlin, Germany: Springer-Verlag.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory & Cognition*, *30*(7), 1138–1147.

Conte, E., Khrennikov, Y. A., Todarello, O., Federici, A., & Zbilut, J. P. (2009). Mental states follow quantum mechanics during perception and cognition of ambiguous figures. *Open*

Systems & Information Dynamics, *16*, 1-17.

Costello, F. J. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, *22*(3), 213–234.

Costello, F. J., & Watts, P. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. *Psychological review*, *121*(3), 463.

Coyle, T. R., Read, L. E., Gaultney, J. F., & Bjorklund, D. F. (1998). Giftedness and variability in strategic processing on a multitrial memory task: Evidence for stability in gifted cognition. *Learning and Individual Differences*, *10*(4), 273–290.

Dennis, M. J., & Ahn, W.-K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition*, *29*(1), 152–164.

Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (p. 249-267). Cambridge: Cambridge University Press.

Eder, A. B., Fiedler, K., & Hamm-Eder, S. (2011). Illusory correlations revisited: The role of pseudocontingencies and working-memory capacity. *The Quarterly Journal of Experimental Psychology*, *64*(3), 517–532.

Eisenberg, P., & Becker, C. A. (1982). Semantic context effects in visual word recognition, sentence processing, and reading: Evidence for semantic strategies. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(5), 739-756.

Elqayam, S., & Evans, J. S. B. T. (2013). Rationality in the new paradigm: strict versus soft bayesian approaches. *Thinking and Reasoning*, *19*, 453-470.

Evans, J. S. B. T., Handley, S. J., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory and Cognition*, *35*(7), 1772–1784.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 678-693.

Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659-676.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25–42.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *Journal of General Psychology*, *113*, 351–357.

Fuss, I. G., & Navarro, D. J. (2013). Open parallel cooperative and competitive decision processes: A potential provenance for quantum probability decision models. *Topics in Cognitive Science*, *5*(4), 818–843.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the bayesians got their beliefs (and what those beliefs actually are): comment on bowers and davis (2012). *Psychological Bulletin*, *138*, 415-422.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334-384.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (p. 86-100). Oxford: Oxford University Press.

Hagmayer, Y., & Waldmann, M. R. (2002). A constraint satisfaction model of causal learning and reasoning. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the cognitive science society* (p. 405-410). Mahwah, NJ: Erlbaum.

Halliwell, J. J. (2014). Two proofs of Fine's theorem. *Physics Letters A*, *378*(40), 2945.

Halliwell, J. J., & Yearsley, J. M. (2013). Negative probabilities, Fine's theorem and linear positivity. *Physical Review A*, *87*, 022114.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*, *101*, 252-254.

Haven, E., & Sozzo, S. (in press). A generalized probability framework to model economic agents' decisions under uncertainty. *International Review of Financial Analysis*.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

JASP Team. (2016). *Jasp.* Retrieved from `https://jasp-stats.org`

Kahneman, D., & Tversky, A. (1972). On prediction and judgment. *ORI Research Monographs*, *12*.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Kim, J. H., & Pearl, J. (1983). A computational model for causal and diagnostic reasoning in inference systems. In *Proceedings of the 8th international joint conference on artificial intelligence (ijcai)* (pp. 190–193).

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1-17.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*(3), 430.

Kutzner, F., Vogel, T., Freytag, P., & Fiedler, K. (2011). A robust classic: Illusory correlations are maintained under extended operant learning. *Experimental Psychology*, *58*, 443-453.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences*, *112*(34), 10645–10650.

Liu, A. Y. (1975). Specific information effect in probability estimation. *Perceptual and Motor Skills*, *41*, 475-478.

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

Medin, D. L., & Bazerman, M. H. (1999). Broadening behavioral decision research: Multiple levels of cognitive processing. *Psychonomic Bulletin & Review*, *6*(4), 533–546.

Meehl, P., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs of patterns, or cutting scores. *Psychological Psychological Bulletin*, *52*, 194-215.

Mistry, P. K., Trueblood, J. S., Vandekerckhove, J., & Pothos, E. M. (submitted). A hierarchical bayesian approach to investigating individual differences in causal reasoning using quantum probability theory.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331-355.

Nielsen, M. A., & Chuang, I. L. (2000). *Quantum computation and quantum information*. Cambridge University Press.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455–485.

Oaksford, M., & Chater, N. (2009). Précis of bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*(01), 69–84.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186–216.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Peres, A. (1998). *Quantum theory: Concepts and methods*. New York: Kluwer Academic.

Plummer, M., et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical* *computing* (Vol. 124, p. 125).

Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J., & Hampton, J. A. (2015). Progress and current challenges with the quantum similarity model. *Frontiers in Psychology*, *6*, 205.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational' decision theory. *Proceedings of the Royal Society B*, *276 (1165)*, 2171-2178.

Pothos, E. M., Busemeyer, J. R., Shiffrin, R. M., & Yearsley, J. M. (in press). The rational status of quantum cognition. *Journal of Experimental Psychology: General*.

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, *120*(3), 679.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science*, *27*, 709-748.

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141-59.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*(3), 323-360.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109–139.

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? markov violations, weak inferences, and failures of explaining away. *Cognitive psychology*, *87*, 88–134.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper

Saddle River, New Jersey: Prentice Hall.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144-1167.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A bayesian hierarchical approach. *Psychological Review*, *120*(1), 39-64.

Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology*, *85*, 181–191.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99–118.

Sloman, S. A., & Fernbach, P. M. (2011). Human representation and reasoning about complex causal systems. *Information, Knowledge, Systems Management*, *10*, 1-15.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*(1), 235.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289.

Trueblood, J. S., & Busemeyer, J. R. (2012). A quantum probability model of causal reasoning. *Frontiers in Cognitive Science*, *3*, 1-13.

Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making* (pp. 141–162). Springer.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293-315f.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550-592.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from bayes's theorem and the additivity principle. *Memory & Cognition*, *30*(2), 171–178.

von Neumann, J. (1932). *Mathematical foundations of quantum mechanics*. Princeton, NJ: Princeton University Press.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind. prospects for bayesian cognitive science* (pp. 453–484). Oxford: Oxford: University Press.

Walker, L., Thibaut, J., & Andreoli, V. (1972). Order of presentation at trial. *Yale Law Journal*, *82*, 216–226.

Wang, Z., & Busemeyer, J. R. (2013). A quantum question order model supported by empirical tests of an a priori and precise prediction. *Topics in Cognitive Science*, *5*, 689-710.

Yearsley, J. M. (in press). Advanced tools and concepts for quantum cognition: A tutorial. *Journal of Mathematical Psychology*, 3.

Yearsley, J. M., & Pothos, E. M. (2016). Zeno's paradox in decision-making. *Proceedings of the Royal Society B*, *238*, 20160291.

Appendix A

Details on Model Parameterization

In this appendix we provide some additional details for how the 2D and 4D models were parameterized.

2D model

We begin by noting that we take the initial state to be a diagonal matrix. We can always do this by a suitable choice of basis, and it turns out to be useful because the conditions under which a general matrix is an allowable density matrix are difficult to express algebraically (the key one is all eigenvalues are non-negative).

Different bases in our space are related by 2D unitary transformations, which we chose to parameterize in the following way

$$R_j = \begin{pmatrix} \cos(\theta_j) & -\sin(\theta_j)e^{i\phi_j} \\ \sin(\theta_j)e^{-i\phi_j} & \cos(\theta_j) \end{pmatrix}. \qquad (35)$$

We have two things to argue, firstly that this form is general enough for our purposes, and secondly that we may in fact take the additional step of setting one of the $\phi_j = 0$.

In general, any two bases in this 2D Hilbert space can be connected by a unitary transformation. Matrices for the form Eq.(35) are not the most general possible unitary transformations, which would also have phase factors on the diagonal elements. However when computing the relevant projection operators, such as $P_{X_1} = R_X P_{E_1} R_X^\dagger$ the resulting expressions depend only on the sum of the phases. Without loss of generality, therefore, we may choose the phase of the diagonal elements to be 0. This proves that the form Eq.(35) is general enough for our needs.

Next we need to argue that we can set $\phi_E = 0$ without loss of generality. To see why this is the case, consider the following argument. We are interested in expressions like $\mathrm{Tr}(P_B P_A \rho P_A)$. Consider the unitary matrix

$$\Phi = \begin{pmatrix} e^{i\phi} & 0 \\ 0 & e^{-i\phi} \end{pmatrix} \tag{36}$$

it is easy to see that the operation $\tilde{M} = \Phi M \Phi^\dagger$ sends

$$\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \to \begin{pmatrix} m_{11} & m_{12}e^{2i\phi} \\ m_{21}e^{-2i\phi} & m_{22} \end{pmatrix}, \tag{37}$$

thus changing only the phase of the off diagonal terms. In particular $\tilde{\rho} = \Phi \rho \Phi^\dagger = \rho$. Since Φ is unitary $\Phi^\dagger \Phi = 1$ and so $\mathrm{Tr}(P_B P_A \rho P_A) = \mathrm{Tr}(\Phi P_B \Phi^\dagger \Phi_A \Phi^\dagger \Phi \rho \Phi^\dagger \Phi P_A \Phi^\dagger \Phi P_B \Phi^\dagger) = \mathrm{Tr}(\tilde{P}_B \tilde{P}_A \rho \tilde{P}_A)$ so that all expressions for probabilities are left unchanged if we add a constant to all the ϕ_i parameters. We may therefore set one of them to zero without changing any of the computed probabilities.

This proves that we may set $\phi_E = 0$ without loss of generality.

4D model

We chose to parameterize the $4\mathrm{D}_{\mathrm{IC}}$ unitary transformations in the following way

$$R = \begin{pmatrix} \cos(\theta_1) & -\sin(\theta_1)e^{i\phi_1} & 0 & 0 \\ \sin(\theta_1)e^{-i\phi_1} & \cos(\theta_1) & 0 & 0 \\ 0 & 0 & \cos(\theta_2) & -\sin(\theta_2)e^{i\phi_2} \\ 0 & 0 & \sin(\theta_2)e^{-i\phi_1} & \cos(\theta_2) \end{pmatrix} \tag{38}$$

where $\theta_1, \theta_2, \phi_1, \phi_2$ are real angles. In a similar way to the 2D model, R is not the most general possible transformation. Essentially we are arguing that R may be written in the form

$$R = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \tag{39}$$

where U_1, U_2 are 2x2 unitary matrices. Therefore exactly as in the 2D model, U_1 and U_2 could have phases attached to the diagonal elements, but any expressions for probabilities would be equivalent to instead changing $\phi_{1/2}$.

We also need to argue that we can take the elements of the matrix S to be real. Start with a more general S where the diagonal elements are not real. The argument proceeds in exactly the same way as for the 2D case, one shows that a transformation exists which serves to change the phase factors of the off-diagonal elements of R and S while leaving ρ and any probabilities invariant. We can therefore choose the elements of the matrix S to be real without loss of generality.

For the $4\mathrm{D}_{\mathrm{CC}}$ model, the arguments for reducing the number of parameters is identical to the $4D_{IC}$ case.

Appendix B

Stimuli from Experiment 1

African Lake Shrimp

Cover Story

Biologists have recently discovered a new kind of shrimp, which are found in the African Great Lakes. The biologists have found identical shrimp in all nine African Great Lakes. They were able to establish three distinctive features of these shrimp.

Features

(F1) ACh neurotransmitter: The shrimp use acetylcholine (ACh) as a brain neurotransmitter. They either have a high or low amount of ACh. Most shrimp have a high amount of ACh whereas a few have a low amount of ACh.

(F2) Sleep cycle: The shrimp have either an accelerated or normal sleep cycle. Most shrimp have a normal sleep cycle (12 hours sleep, 12 hours awake) whereas a few have an accelerated sleep cycle (4 hours sleep, 4 hours awake).

(F3) Body weight: The shrimp either have a high or low body weight. Half of the shrimp have a high body weight whereas half have a low body weight.

Causal Relationships

(F1 → F3). A high quantity of ACh neurotransmitter causes a high body weight. The neurotransmitter stimulates greater feeding behavior, which results in more food ingestion and more body weight. A low quantity of ACh neurotransmitter causes a low body weight. Low quantities of the neurotransmitter stimulate less feeding behavior, which result in less food ingestion and body weight.

(F2 → F3). An accelerated sleep cycle causes a high body weight. Shrimp habitually feed after waking, and shrimp on an accelerated sleep cycle wake three times a day instead of once. A normal sleep cycle causes a low body weight. Shrimp on a normal sleep cycle wake once a day and thus feed once a day.

Kehoe Ants

Cover Story

Biologists have recently discovered a new kind of ant, which are found in the forests of the volcanic island of Kehoe. The biologists have found identical ants in all nine Kehoe National Forests. They were able to establish three distinctive features of these ants.

Features

(F1) Iron sulfate: The ants have blood that contains iron sulfate. They either have a high or low amount of iron sulfate in their blood. Most ants have a high amount of iron sulfate whereas a few have a low amount of iron sulfate.

(F2) Immune system: The ants have either a hyperactive or normal immune system. Most ants have a normal immune system whereas a few have a hyperactive immune system.

(F3) Blood: The ants either have thick or thin blood. Half of the ants have thick blood whereas half have thin blood.

Causal Relationships

(F1 → F3). A high quantity of iron sulfate causes thick blood. Iron sulfate provides the extra iron that the ants use to produce extra red blood cells. The extra red blood cells thicken the blood. A low quantity of iron sulfate causes thin blood. Low quantities of iron sulfate result in lower levels of red blood cells, which result in thin blood.

(F2 → F3). A hyperactive immune system causes thick blood. A hyperactive immune system accelerates the production of blood cells, which thickens the blood. A normal immune system causes thin blood. Ants with a normal immune system produce fewer blood cells and thus have thin blood.

Appendix C

Parameter estimates from Experiment 1

The mean and 95% highest density intervals (HDIs) for the parameter estimates of each of the five models in Experiment 1 are given in Tables C1 and C2.

Author Note

JST and JMY were supported by NSF grant SES-1556415. EMP was supported by Leverhulme Trust grant RPG-2015-311 and H2020-MSCA-IF-2015 grant 696331. Parts of this research were reported in the Proceedings of the 38th Annual Conference of the Cognitive Science Society. All authors contributed in a significant way to the manuscript and all authors approved the final manuscript. The authors did not have any conflicts of interest that would have impacted this work.

Footnotes

[1] A Hilbert space is a generalization of the notion of a Euclidean space (indeed a Euclidean space *is* a Hilbert space). A Hilbert space is a vector space, equipped with an inner product which has some convergence properties. The spaces we will use in this paper are simply 2, 4 and 8D complex vector spaces where the inner product is taken to be the usual dot product between vectors.

[2] It is possible that participants who wanted to respond “don’t know” were attracted to the “equally likely” option more than the other two options. However, we think that this is unlikely in our task. Participants were never instructed to use the “equally likely” response option as a “don’t know” response. We included an “equally likely” option because participants were instructed that the probability of the effect (e.g., high or low body weight of a shrimp) was 0.5. Thus, it is very reasonable to assume this response maps to the probability 0.5 for data analyses.

[3] An interesting question is whether we can give a direct interpretation to the various parameters in a given quantum model. It is clear that in some cases, such as for ρ or ε in the 2D POVM model, a reasonably simple interpretation can be given. However interpreting the values of the various angles that appear in the models is more problematic. First the observed probabilities typically depend on various combinations of angles, so that changing the value of one angle can be offset by varying another. Second, while the size of the angle between events determines the size of any order effects (via the commutator,) it is less clear that we can use this to define something like a ‘degree’ of incompatibility. An important future topic for research is to understand how to interpret best fit values from quantum models.

[4] BF_M is the Bayes factor comparing the specified model against all other possible models.

Table 1: Summary of the models and their properties

Model	Number of Parameters	Properties					
		Incompatible Events	Order Effects	Reciprocity	Memorylessness	Markov violations	Anti-discounting
2D	6	E,X,Y	$X,Y;E,X;E,Y$	Yes	Yes	Yes	Yes
2D POVM	7	E,X,Y	$X,Y;E,X;E,Y$	No**	No**	Yes	Yes
4D_{IC}	10*	X,Y	X,Y	No	No	Yes	Yes
4D_{CC}	10*	$E,X;E,Y$	$E,X;E,Y$	No	No	Maybe†	Maybe†
8D	8*	None	None	No	No	Maybe†	Maybe†

*The normalization constraint means that the degrees of freedom in the model is one less than the number of parameters.

**The 2D POVM model predicts these properties when $\varepsilon \to 0$

†This is dependent on the state vector

Table 2: Summary of Experiments

	Experiment 1	Experiment 2	Experiment 3
Effects	order effects “cause-effect” reciprocity “cause-cause” memorylessness	both types of reciprocity “cause-effect” memorylessness Markov violations anti-discounting	order effects “cause-cause” reciprocity “cause-cause” memorylessness Markov violations
Response Format	ternary choice	rating from 0 to 100	ternary choice
Other Features	verbal base-rates	verbal base-rates	numeric base-rates, CRT extended testing (6 blocks)
Models	2D, 2D POVM, 4D$_{\text{IC}}$, 4D$_{\text{CC}}$, 8D, 8D causal power	2D POVM, 4D$_{\text{CC}}$, 8D	2D POVM, 4D$_{\text{CC}}$, 8D
Analyses	group-level and hierarchical modeling* cross validation	individual-level modeling	group-level modeling**

*Results for the hierarchical modeling are provided in the online supplementary material

**Modeling is performed on sub-groups determined by the Cognitive Reflection Test

Table 3: Questions in Experiment 1 along with mean choice scores.

Block	Elementary events		Conditionals with one event		Conditionals with two events	
	Event	Mean	Event	Mean	Event	Mean
BX	E	0.50 (0.09)	$X\|E_1$	0.96 (0.17)	$E\|X_1,Y_1$	0.95 (0.20)
	X	0.92 (0.21)	$Y\|E_1$	0.80 (0.37)	$E\|X_1,Y_2$	0.42 (0.32)
	Y	0.17 (0.32)	$E\|X_1*$	0.94 (0.16)	$E\|X_2,Y_1$	0.63 (0.30)
			$E\|X_2*$	0.05 (0.14)	$E\|X_2,Y_2$	0.07 (0.17)
BY	E	0.52 (0.13)	$X\|E_2$	0.10 (0.29)	$E\|Y_1,X_1$	0.93 (0.22)
	X	0.85 (0.32)	$Y\|E_2$	0.09 (0.24)	$E\|Y_1,X_2$	0.37 (0.30)
	Y	0.15 (0.31)	$E\|Y_1*$	0.91 (0.23)	$E\|Y_2,X_1$	0.68 (0.28)
			$E\|Y_2*$	0.10 (0.22)	$E\|Y_2,X_2$	0.08 (0.23)

*Events that participants saw twice within the same block.

Responses were scored by assigning the following values to the three response options: feature value 1 = 1, feature value 2 = 0, and equally likely = 0.5. Standard deviations are given in parentheses.

Table 4: Bayesian paired samples t-tests for order effects, reciprocity, and memorylessness in Experiment 1.

Effect	Comparison	BF_{10}
Order effects	$E\|X_1Y_1$ and $E\|Y_1X_1$	0.374
	$E\|X_1Y_2$ and $E\|Y_2X_1$	172.525
	$E\|X_2Y_1$ and $E\|Y_1X_2$	90.012
	$E\|X_2Y_2$ and $E\|Y_2X_2$	0.150
Reciprocity	$X\|E_1$ and $E\|X_1$	0.201
	$X\|E_2$ and $E\|X_2$	0.303
	$Y\|E_1$ and $E\|Y_1$	1.103
	$Y\|E_2$ and $E\|Y_2$	0.149
Memorylessness	$E\|X_1$ and $E\|Y_1,X_1$	0.145
	$E\|X_1$ and $E\|Y_2,X_1$	1.418e+7
	$E\|X_2$ and $E\|Y_1,X_2$	2.231e+8
	$E\|X_2$ and $E\|Y_2,X_2$	0.220
	$E\|Y_1$ and $E\|X_1,Y_1$	0.594
	$E\|Y_1$ and $E\|X_2,Y_1$	364018.502
	$E\|Y_2$ and $E\|X_1,Y_2$	4.192e+7
	$E\|Y_2$ and $E\|X_2,Y_2$	0.252

Table 5: Cross validation results for Experiment 1.

Model	DIC for first group	MSE for second group
2D	-19.70	0.022
2D POVM	-37.59	0.010
$4D_{IC}$	-1.94	0.070
$4D_{CC}$	-40.32	0.010
8D	-24.79	0.018
8D causal power	-26.75	0.021

Table 6: Conditional judgment questions in Experiment 2 along with mean judgments.

Conditionals with causes		Conditionals with effects and causes		Conditionals with effects and two causes	
Event	Mean	Event	Mean	Event	Mean
$X_1\|Y_1$	59.01 (22.28)	$E_1\|X_1$	64.92 (18.16)	$X_1\|E_1,Y_1$	62.42 (22.32)
$X_1\|Y_2$	39.71 (23.17)	$E_1\|Y_1$	66.20 (20.57)	$X_1\|E_1,Y_2$	44.39 (22.79)
$Y_1\|X_1$	60.39 (24.01)	$X_1\|E_1$*	65.82 (16.56)	$Y_1\|E_1,X_1$	64.92 (21.57)
$Y_1\|X_2$	41.16 (21.40)	$Y_1\|E_1$*	68.54 (17.02)	$Y_1\|E_1,X_2$	42.97 (21.67)

*Events that participants judged twice within the same block.
Standard deviations are given in parentheses.

Table 7: Bayesian paired samples t-tests for reciprocity, memorylessness, Markov violations, and anti-discounting in Experiment 2.

Effect	Comparison	BF_{10}
Reciprocity	$X_1\|E_1$ and $E_1\|X_1$	0.157
	$Y_1\|E_1$ and $E_1\|Y_1$	0.283
	$X_1\|Y_1$ and $Y_1\|X_1$	0.158
Memorylessness	$X_1\|Y_1$ and $X_1\|E_1,Y_1$	0.232
	$X_1\|Y_2$ and $X_1\|E_1,Y_2$	0.599
	$Y_1\|X_1$ and $Y_1\|E_1,X_1$	0.559
	$Y_1\|X_2$ and $Y_1\|E_1,X_2$	0.176
Markov violations	$X_1\|Y_1$ and $X_1\|Y_2$	942.4
	$Y_1\|X_1$ and $Y_1\|X_2$	261.3
Anti-discounting	$X_1\|E_1$ and $X_1\|E_1,Y_1$	0.255
	$Y_1\|E_1$ and $Y_1\|E_1,X_1$	0.335

Table 8: Bayesian Pearson correlations between effects in Experiment 2.

		MemorylessnessScore	MarkovScore	AntidiscountingScore
ReciprocityScore	Person's r	0.397	-0.328	0.054
	BF_{10}	16.900	3.551	0.177
MemorylessnessScore	Person's r	-	-0.331	-0.351
	BF_{10}	-	3.768	5.717
MarkovScore	Person's r	-	-	-0.195
	BF_{10}	-	-	0.469

Table 9: Questions in Experiment 3 along with mean choice scores.

Block	Elementary events		Conditionals with one event		Conditionals with two events	
	Event	Mean	Event	Mean	Event	Mean
BX_i	E	0.52 (0.07)	$E\|X_1$	0.90 (0.19)	$E\|X_1,Y_2$	0.40 (0.23)
	X	0.93 (0.14)	$E\|X_2$	0.08 (0.16)	$E\|X_2,Y_1$	0.67 (0.22)
	Y	0.11 (0.22)	$X\|Y_2$	0.62 (0.35)		
			$Y\|X_1$	0.38 (0.35)		
BY_i	E	0.53 (0.11)	$E\|Y_1$	0.91 (0.19)	$E\|Y_1,X_2$	0.36 (0.20)
	X	0.93 (0.14)	$E\|Y_2$	0.13 (0.22)	$E\|Y_2,X_1$	0.66 (0.21)
	Y	0.16 (0.26)	$X\|Y_1$	0.77 (0.25)		
			$Y\|X_2$	0.25 (0.28)		

Responses were scored by assigning the following values to the three response options: feature value 1 = 1, feature value 2 = 0, and equally likely = 0.5. Standard deviations are given in parentheses.

Table 10: Bayesian paired samples t-tests for order effects, reciprocity, memorylessness, and Markov violations in Experiment 3 for first, middle, and last blocks.

Effect	Comparison	BF_{10} First Blocks	BF_{10} Middle Blocks	BF_{10} Last Blocks
Order effects	$E\|X_1Y_2$ and $E\|Y_2X_1$	213351.78	126.86	92.09
	$E\|X_2Y_1$ and $E\|Y_1X_2$	673868.48	3551.65	425.53
Reciprocity	$X\|Y_1$ and $Y\|X_1$	40940.84	349.37	212562.76
	$X\|Y_2$ and $Y\|X_2$	84599.69	63858.31	3364.80
Memorylessness	$E\|X_1$ and $E\|Y_2,X_1$	534.23	90371.02	849770.96
	$E\|X_2$ and $E\|Y_1,X_2$	5.229e+6	3.147e+7	3.684e+9
	$E\|Y_1$ and $E\|X_2,Y_1$	319.60	1.264e+7	10446.62
	$E\|Y_2$ and $E\|X_1,Y_2$	73307.93	20655.73	849770.96
Markov violations	$X\|Y_1$ and $X\|Y_2$	14.3	0.34	1.47
	$Y\|X_1$ and $Y\|X_2$	1.14	1.73	0.30

Table C1: Best fit model parameters and HDIs for each of the models fit in Experiment 1.

Model	Parameter	Mean	95% HDI
2D	ρ	0.586	$[0.500, 0.693]$
	θ_E	0.768	$[0.471, 1.063]$
	θ_X	0.344	$[0.006, 0.783]$
	ϕ_X	1.178	$[0.468, 1.563]$
	θ_Y	1.194	$[0.810, 1.558]$
	ϕ_Y	1.199	$[0.477, 1.561]$
	λ	10.744	$[3.939, 17.880]$
	τ	0.282	$[0.135, 0.429]$
2D POVM	ρ	0.578	$[0.534, 0.626]$
	θ_E	0.782	$[0.705, 0.857]$
	θ_X	0.310	$[0.105, 0.528]$
	ϕ_X	1.270	$[0.835, 1.560]$
	θ_Y	1.171	$[0.956, 1.393]$
	ϕ_Y	1.367	$[1.096, 1.562]$
	ε	0.032	$[0.011, 0.050]$
	λ	28.596	$[10.910, 48.986]$
	τ	0.290	$[0.152, 0.426]$
$4D_{CC}$	ρ_{11}	0.354	$[0.244, 0.456]$
	ρ_{22}	0.179	$[0.022, 0.296]$
	ρ_{33}	0.071	$[0.000, 0.184]$
	$\rho_{44}*$	0.396	$[0.289, 0.492]$
	θ_1	0.707	$[0.657, 0.756]$
	θ_2	0.781	$[0.770, 0.791]$
	ϕ_1	0.791	$[0.147, 1.417]$
	ϕ_2	0.996	$[0.309, 1.541]$
	θ_a	0.802	$[0.242, 1.286]$
	θ_b	0.429	$[0.025, 1.122]$
	λ	34.792	$[11.503, 60.631]$
	τ	0.224	$[0.095, 0.357]$

*The normalization constraint means that the degrees of freedom in the model is one less than the number of parameters.

Table C2: Best fit model parameters and HDIs for each of the models fit in Experiment 1 continued.

Model	Parameter	Mean	95% HDI
$4D_{IC}$	ρ_{11}	0.256	$[0.160, 0.347]$
	ρ_{22}	0.245	$[0.156, 0.343]$
	ρ_{33}	0.211	$[0.142, 0.277]$
	$\rho_{44}*$	0.287	$[0.222, 0.357]$
	θ_1	0.672	$[0.229, 1.136]$
	θ_2	0.670	$[0.228, 1.139]$
	ϕ_1	0.902	$[0.241, 1.512]$
	ϕ_2	0.782	$[0.154, 1.397]$
	θ_a	0.746	$[0.112, 1.465]$
	θ_b	0.638	$[0.126, 1.073]$
	λ	3.503	$[2.000, 5.445]$
	τ	0.246	$[0.105, 0.394]$
8D	ρ_{11}	0.196	$[0.137, 0.252]$
	ρ_{22}	0.131	$[0.084, 0.180]$
	ρ_{33}	0.104	$[0.054, 0.153]$
	ρ_{44}	0.100	$[0.052, 0.147]$
	ρ_{55}	0.071	$[0.021, 0.123]$
	ρ_{66}	0.072	$[0.020, 0.124]$
	ρ_{77}	0.130	$[0.081, 0.180]$
	$\rho_{88}*$	0.196	$[0.137, 0.252]$
	λ	18.165	$[5.859, 32.721]$
	τ	0.266	$[0.133, 0.404]$
8D causal power	$p(X_1)$	0.528	$[0.503, 0.557]$
	$p(Y_1)$	0.473	$[0.444, 0.498]$
	w_X	0.213	$[0.104, 0.321]$
	w_Y	0.191	$[0.090, 0.294]$
	w_a	0.378	$[0.299, 0.449]$
	λ	14.276	$[5.080, 24.874]$
	τ	0.265	$[0.127, 0.407]$

*The normalization constraint means that the degrees of freedom in the model is one less than the number of parameters.

Figure Captions

Figure 1. The posterior distributions for the predictions of the six models plotted against the data from Experiment 1. The (red) circles are the average choice scores for the different questions in the experiment. The black squares show the posterior mass for the model predictions. In each plot, questions are ordered with respect to their mean choice scores from smallest to largest.

Figure 2. Scatter plots for the four effects examined in Experiment 2. Each point represents an individual-level judgment. The dotted black line is the 45 degree line of identity. The solid black line is the best fit line for a model predicting the second judgment from the first assuming an intercept of 0. The beta weight is included for each effect. The top left panel shows the three different pairs of judgments for reciprocity: red diamond represents the judgments $X_1|E_1$ (horizontal axis) and $E_1|X_1$ (vertical axis), blue square represents the judgments $Y_1|E_1$ and $E_1|Y_1$, and green circle represents the judgments $X_1|Y_1$ and $Y_1|X_1$. The top right panel shows the four different paris of judgments for memorylessness: red diamond represents the judgments $X_1|Y_1$ and $X_1|E_1,Y_1$, blue square represents the judgments $X_1|Y_2$ and $X_1|E_1,Y_2$, green circle represents the judgments $Y_1|X_1$ and $Y_1|E_1,X_1$, and purple asterisk represents the judgments $Y_1|X_2$ and $Y_1|E_1,X_2$. The bottom left panel shows the two different pairs of judgments for Markov violations: red diamond represents the judgments $X_1|Y_1$ and $X_1|Y_2$, and blue square represents the judgments $Y_1|X_1$ and $Y_1|X_2$. The bottom right panel shows the two different pairs of judgments for anti-discounting: red diamond represents the judgments $X_1|E_1$ and $X_1|E_1,Y_1$, and blue square represents the judgments $Y_1|E_1$ and $Y_1|E_1,X_1$.

Figure 3. Behavioral results for each of the subgroups of participants identified as being best fit by the 2D POVM, $4D_{CC}$ and 8D models in Experiment 2. Low ReciprocityScores, MemorylessnessScores, and AntidiscountingScores indicate stronger evidence for reciprocity, memorylessness, and anti-discounting respectively. High MarkovScores indicate larger Markov violations. Error bars show the standard error.

Figure 4. Observed probability judgments compared to model predictions from three models (top: 2D POVM, middle: 4D$_{\mathrm{CC}}$, bottom: 8D) for Experiment 2. The probabilities have been split up into three types: probabilities of the form $p(X|Y)$ are shown by (blue) circles, probabilities of the form $p(E|X)$ or $p(X|E)$ are shown as (red) squares, and probabilities of the form $p(E|X,Y)$ or $p(X|E,Y)$ are shown by (green) triangles.

Figure 5. Behavioral results from Experiment 3. Top left: OrderScore for three CRT groups (low, medium, and high) across block pairs. Top right: ReciprocityScore for three CRT groups across block pairs. Bottom left: MemorylessnessScore for three CRT groups across block pairs. Bottom right: MarkovScore for three CRT groups across block pairs. Error bars show the standard error.

Figure 6. The DIC scores for each of the 2D POVM (solid lines), 4D$_{\mathrm{CC}}$ (dashed lines) and 8D (dotted lines) models as a function of block number for each of the three CRT groups in Experiment 3. Left: Low CRT Group. The 2D POVM is clearly superior in the first blocks, but this vanishes by the end of the experiment. Middle: Medium CRT Group. The 2D POVM model generally performs better here, with no obvious change over blocks. Right: High CRT Group. The 4D$_{\mathrm{CC}}$ and 8D models perform best here, and there is little change over the blocks.

Figure 7. The posterior distributions for the predictions of three models (left: 2D POVM, middle: 4D$_{\mathrm{CC}}$, right: 8D) plotted against the data from selected conditions from Experiment 3. The (red) circles are the average choice scores for the different questions in the experiment. The black squares show the posterior mass for the model predictions. In each plot, questions are ordered with respect to their mean choice scores from smallest to largest.

Figure 1

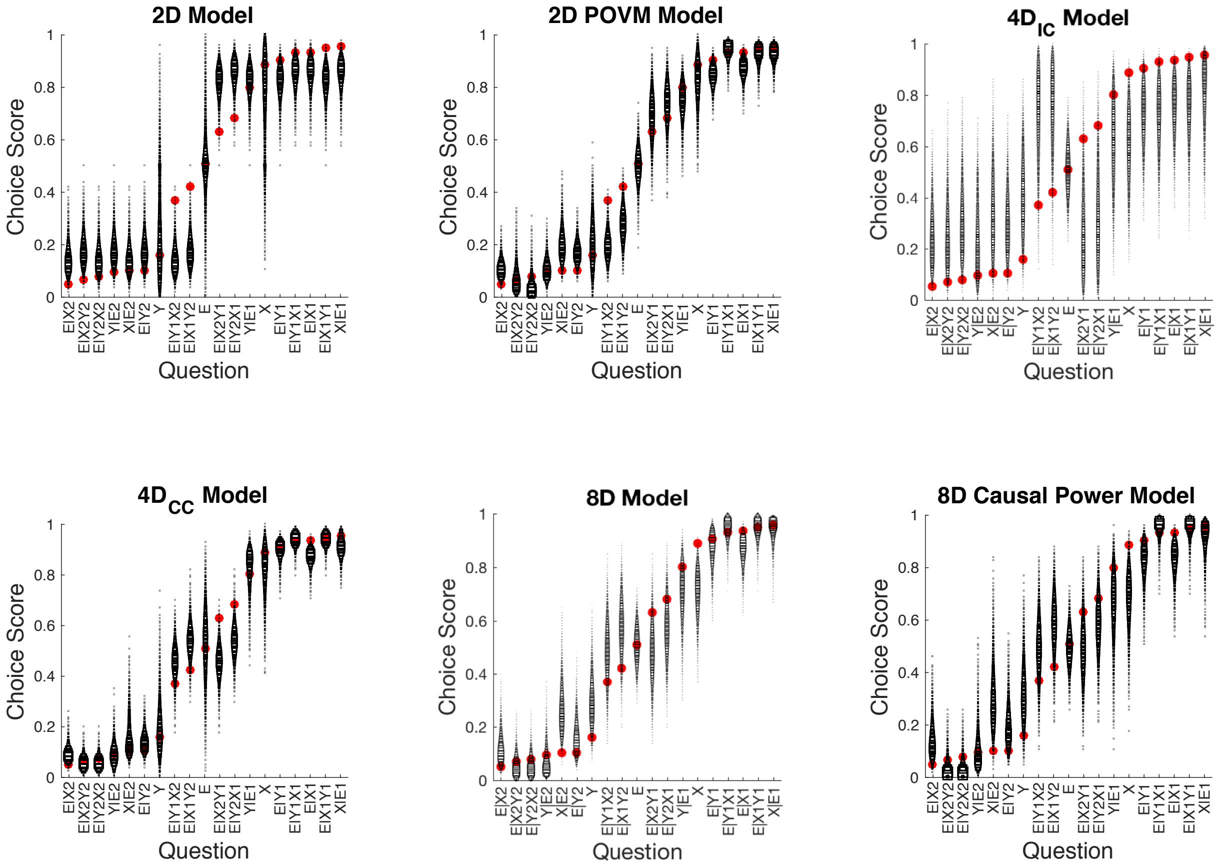

2D Model
2D POVM Model
4D$_{IC}$ Model
4D$_{CC}$ Model
8D Model
8D Causal Power Model
Choice Score
Question
0
0.2
0.4
0.6
0.8
1
EIX2
EIX2Y2
EIY2X2
YIE2
XIE2
EIY2
Y
EIY1X2
EIX1Y2
E
EIX2Y1
EIY2X1
YIE1
X
EIY1
EIY1X1
EIX1
EIX1Y1
XIE1
E|X2
E|X2Y2
E|Y2X2
Y|E2
X|E2
E|Y2
E|Y1X2
E|X1Y2
E|X2Y1
E|Y2X1
Y|E1
E|Y1
E|Y1X1
E|X1
E|X1Y1
X|E1

Figure 2

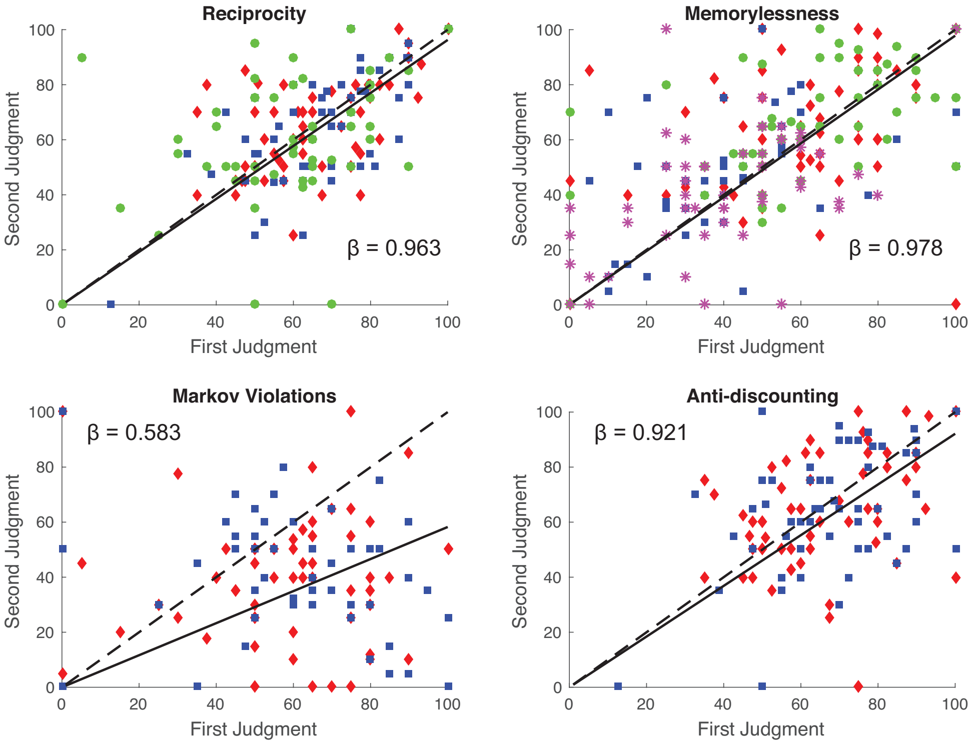

Reciprocity
β = 0.963
Memorylessness
β = 0.978
Markov Violations
β = 0.583
Anti-discounting
β = 0.921
Second Judgment
First Judgment
0
20
40
60
80
100

Figure 3

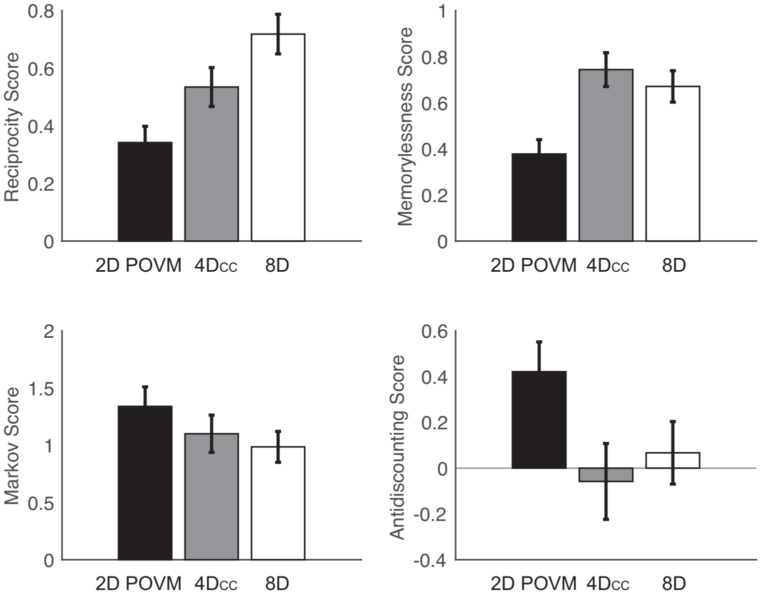
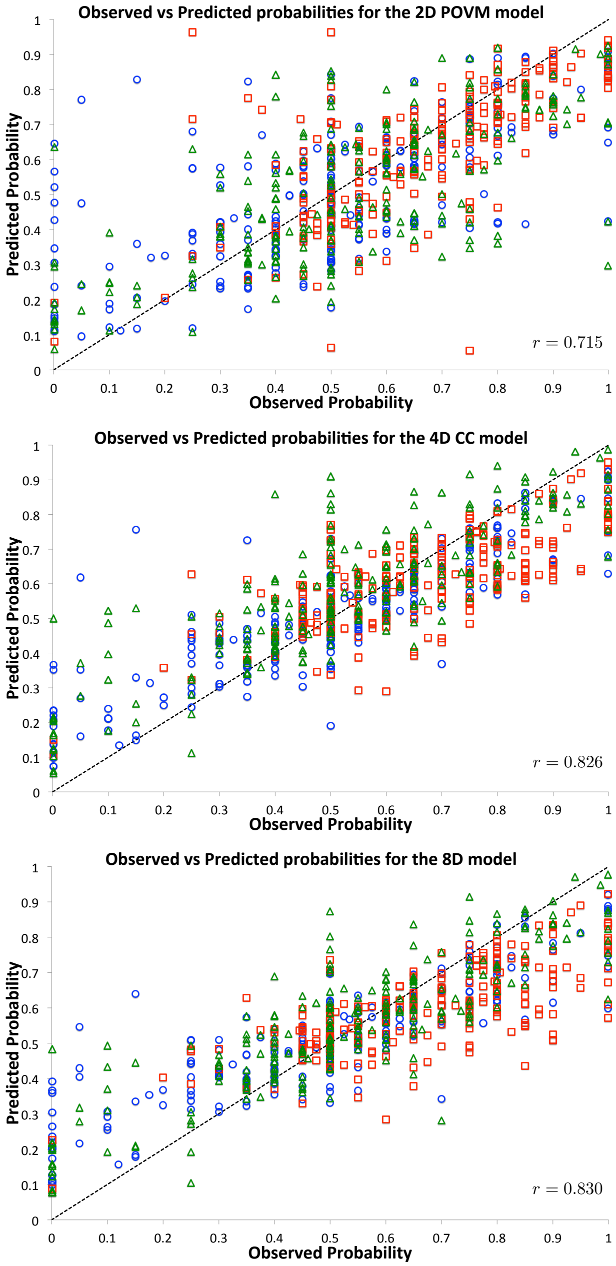

Reciprocity Score
0.8
0.6
0.4
0.2
0
2D POVM
4Dcc
8D
Memorylessness Score
1
0.8
0.6
0.4
0.2
0
2D POVM
4Dcc
8D
Markov Score
2
1.5
1
0.5
0
2D POVM
4Dcc
8D
Antidiscounting Score
0.6
0.4
0.2
0
-0.2
-0.4
2D POVM
4Dcc
8D

Figure 4



Observed vs Predicted probabilities for the 2D POVM model

Observed vs Predicted probabilities for the 4D CC model

Observed vs Predicted probabilities for the 8D model

Figure 5

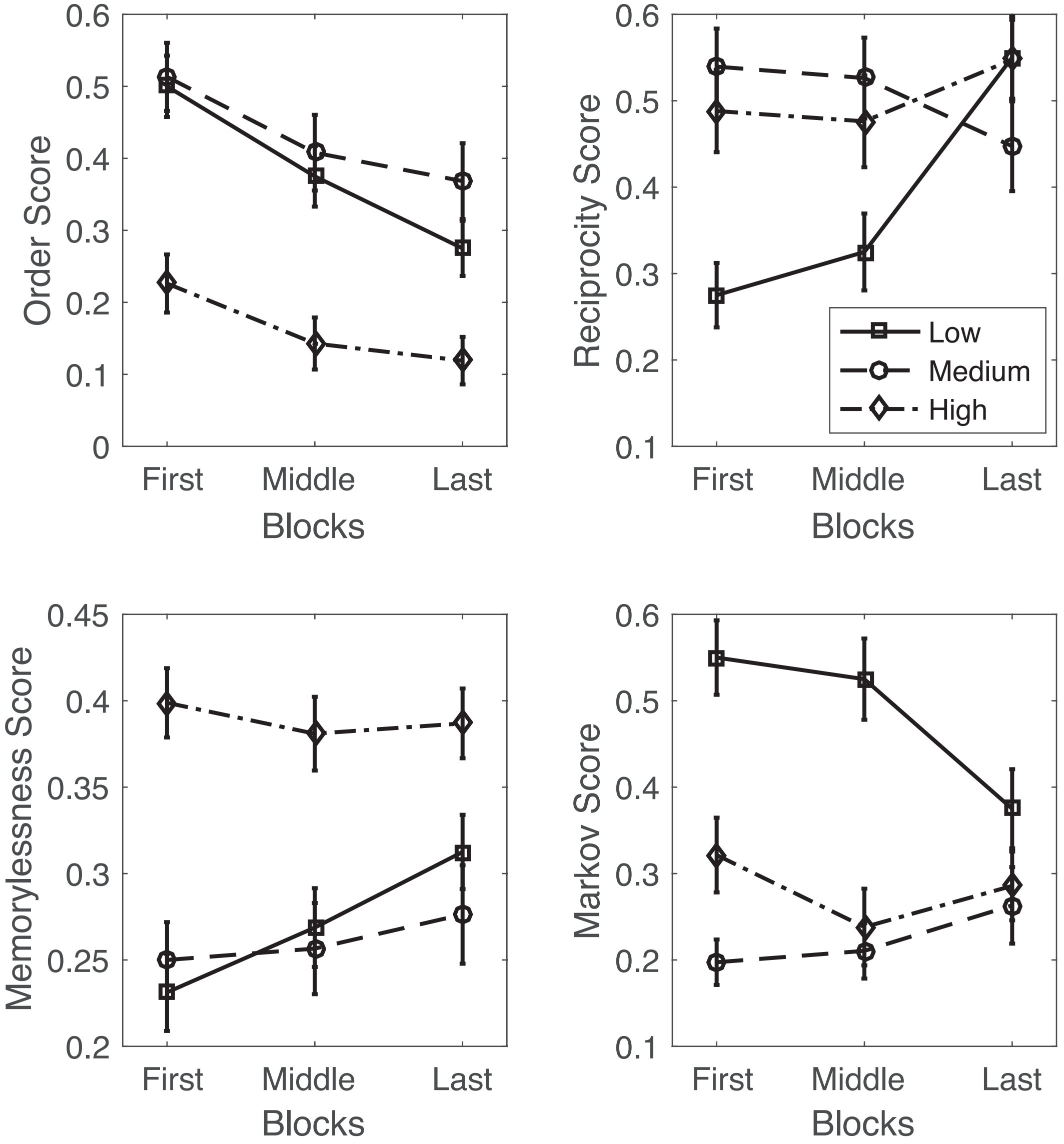

Order Score
0
0.1
0.2
0.3
0.4
0.5
0.6
First
Middle
Last
Blocks
Reciprocity Score
0.1
0.2
0.3
0.4
0.5
0.6
Low
Medium
High
First
Middle
Last
Blocks
Memorylessness Score
0.2
0.25
0.3
0.35
0.4
0.45
First
Middle
Last
Blocks
Markov Score
0.1
0.2
0.3
0.4
0.5
0.6
First
Middle
Last
Blocks

Figure 6

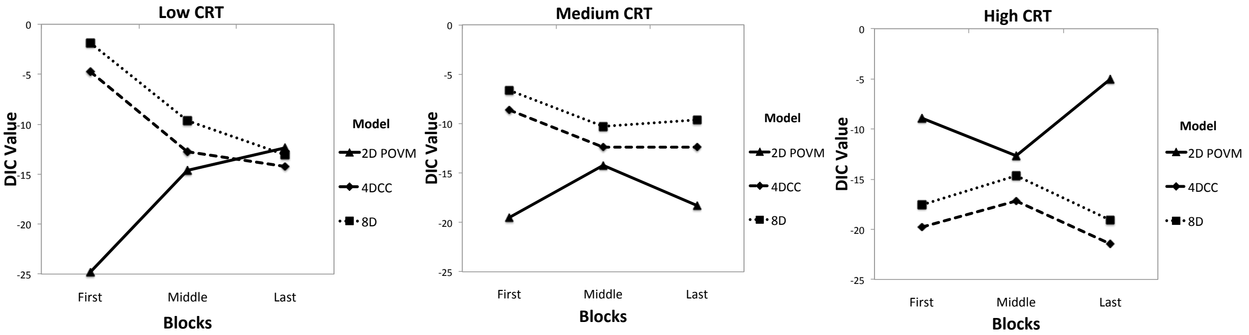
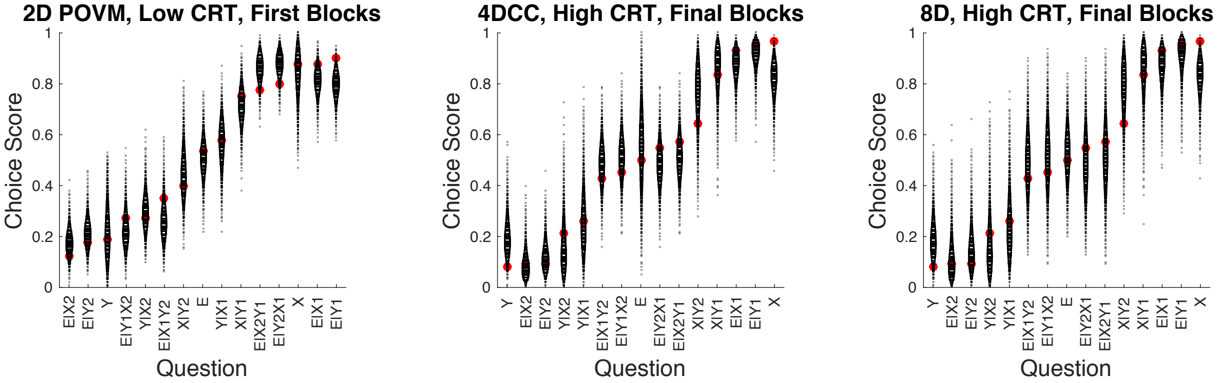

Low CRT
0
-5
-10
-15
-20
-25
DIC Value
First
Middle
Last
Blocks
Model
2D POVM
4DCC
8D
Medium CRT
0
-5
-10
-15
-20
-25
DIC Value
First
Middle
Last
Blocks
Model
2D POVM
4DCC
8D
High CRT
0
-5
-10
-15
-20
-25
DIC Value
First
Middle
Last
Blocks
Model
2D POVM
4DCC
8D

Figure 7


2D POVM, Low CRT, First Blocks
Choice Score
0
0.2
0.4
0.6
0.8
1
E|X2
E|Y2
Y
E|Y1X2
Y|X2
E|X1Y2
X|Y2
E
Y|X1
X|Y1
E|X2Y1
E|Y2X1
X
E|X1
E|Y1
Question
4DCC, High CRT, Final Blocks
Choice Score
0
0.2
0.4
0.6
0.8
1
Y
E|X2
E|Y2
Y|X2
Y|X1
E|X1Y2
E|Y1X2
E
E|Y2X1
E|X2Y1
X|Y2
X|Y1
E|X1
E|Y1
X
Question
8D, High CRT, Final Blocks
Choice Score
0
0.2
0.4
0.6
0.8
1
Y
E|X2
E|Y2
Y|X2
Y|X1
E|X1Y2
E|Y1X2
E
E|Y2X1
E|X2Y1
X|Y2
X|Y1
E|X1
E|Y1
X
Question