Formal Constraints on Cognitive Interpretations of Causal Structure

Jacob Feldman*
Center for Cognitive Science
Rutgers University

Abstract

Human observers are remarkably proficient at extracting the causal structure of both natural and artifactual worlds on the basis of extremely impoverished observations, but a rigorous theory of how they achieve this is elusive. This paper investigates the formal structure of this problem, collapsing the distinction between human and automatic inference about complex systems, and considering an abstract observer in an abstract world. We introduce a *structure algebra*, a reduced description logic that is designed to be biased towards the recovery of causal structure in much the same way as are human observers. We illustrate how the algebra captures human intuitions about the "natural" interpretation of certain canonic types of observed property covariation. Finally we propose a formal criterion for the adequacy of a given description language to capture algebraically the "true" causal structure of a particular closed world.

*Please send correspondence to the author at the Rutgers Center for Cognitive Science (RuCCS), Psychology Building, Busch Campus, Rutgers University, New Brunswick, New Jersey, 08903; or by e-mail at jacob@ruccs.rutgers.edu.

1 Introduction

Human observers are the gold standard for the interpretation of complex systems with many interacting variables. Our environment itself, the world outside our sense organs, is such a system: filled with a myriad of objects and forces that are in principle arbitrarily complicated, and inductively ambiguous to an extreme degree. Effortlessly, though, we make "sense" of all of it, extracting a connecting tissue of underlying causal structure [21, 13]. Our success in so doing is presumably enabled by unconscious interpretive machinery that may well be opaque to introspection. Yet this machinery may well be amenable to mathematical characterization.

To researchers in fields concerned with automatic interpretation, human performance of these interpretive feats is a tantalizing existence proof that such interpretation is possible. But the mere existence of opaque subjective interpretations is not sufficient to aid theory-building. For human cognitive processes to be more tangibly useful in theory-building, a rigorous and complete characterization of them in mechanistic terms is required. Historically, such an understanding has not been available.

This paper presents the outlines of an account of how certain types of complex systems

might be interpreted by human observers, couched in a formal language that collapses the distinction between human and computer inference. The general outlook is one in which the observer makes strong implicit assumptions about the underlying form of causal forces in its environment. This in turn allows strong generalizations from very limited observations—generalizations which tend to be true only in the case that the environment is actually structured in the assumed way. The highlight will be a very general minimal requirement (the Projection Condition) that, when it holds, suffices to guarantee that the observer can successfully infer the "essential" component of the structure actually in effect in its environment.

2 An algebra for the description of observed structure

Consider an observer confronted with a "world" $\mathcal{W}$ consisting of the five amoeba-like objects shown in Fig. 1. (We might just as well consider a world containing events instead of objects—in fact any entities that can be described by logical predicates.)

First, the observer must choose some "language" in which to express the structure of this world, which here we think of simply as a list $\Sigma = \{\sigma_1, \sigma_2, \ldots\}$ of abstract property tags, called the *property set*. For the given world, an appropriate set might be

$$\Sigma = \{\texttt{blob\_shaped}, \texttt{shaded}, \texttt{large}, \texttt{has\_nucleus}, \texttt{has\_dotted\_membrane}\}, \quad (1)$$

which we abbreviate to $\Sigma = \{A, B, C, D, E\}$ under the assignment

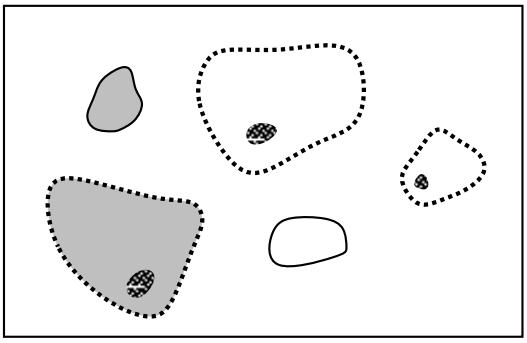


Figure 1: A "world" $\mathcal{W}$ containing amoeba-like objects.

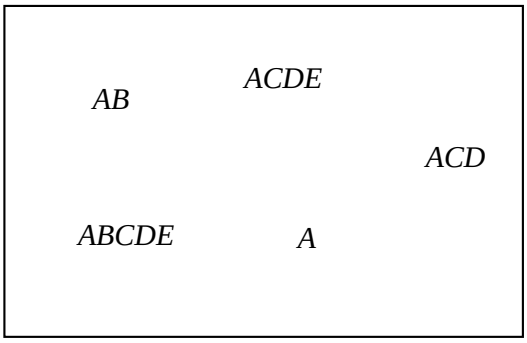


Figure 2: The same world, encoded in the property set Σ.

$$\begin{aligned} A &= \texttt{blob\_shaped}, \\ B &= \texttt{shaded}, \\ C &= \texttt{has\_nucleus}, \\ D &= \texttt{has\_dotted\_membrane}, \\ E &= \texttt{large}. \end{aligned}$$

The same world $\mathcal{W}$ encoded into Σ is shown in Fig. 2. This is really just a set of strings, a subset of 2^{Σ}, which we think of as the world $\mathcal{W}$ "taken at" or "projected into" Σ, i.e. here

$$\mathcal{W}|_\Sigma = \{A, AB, ACD, ACDE, ABCDE\}. \quad (2)$$

This notation explicitly recognizes the dichotomy between the abstract world $\mathcal{W}$ and the concrete string of properties $\mathcal{W}|_\Sigma$ that comes into existence only when the $\mathcal{W}$ is "sampled" at a particular Σ. In particular it allows us to consider the same world projected into alternative property languages, an important consideration below. The set of observations $\mathcal{W}|_\Sigma$ contains some patterns of covariation, with some properties appearing or failing to appear in the company of others. Such patterns are of course not arbitrary, but reflect (in some possibly indirect way) the patterns of causal interaction among the distal properties represented proximally by the various σ_i, which we will refer to loosely as the "causal structure" of the world $\mathcal{W}$ (cf [17]). (A stricter definition will be developed below.)

The goal of the observer is to infer these distal patterns from the extremely impoverished evidence available among the observable patterns. Its leverage in doing so consists of contingent hypotheses about the regularities in force within the closed world in which it finds itself [19], and about how this structure maps to observable patterns (see [11, 23, 24, 20]). Hence our goal as theorists is to characterize the relationship between these distal patterns and the observable ones: literally, the semiotics of the environment.

The set of observations $\mathcal{W}|_\Sigma$ can be thought of as a Boolean logical expression

$$A \vee AB \vee ACD \vee ACDE \vee ABCDE, \quad (3)$$

reading $\vee$ as "or." This expression is a logical "theory" that is satisfied by the environment. (In logical terms the environment is a "model" of this expression.) The Boolean language from which it is built (containing $\vee$, $\wedge$, $\neg$, etc.) is powerful enough to express any possible additional observations that might occur.

Paradoxically, though, this very expressive power is a handicap from the point of view of inferring distal structure. The observations here are inductively ambiguous to an extreme degree; the five observed objects do not *logically* constrain any future observations to manifest any particular structure. Intriguingly, this ambiguity is inherent in the expressive power of the Boolean expression encoding it. That this language is capable of expressing *any* other observations is tantamout to allowing that *any* other observations are possible: a total absence of constraint on induction. This epitomizes a natural trade-off between expressivity and inferential power. By contrast, a *weaker* logic, suitably chosen, would allow a *stronger* inference of structure.

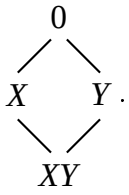
A weaker logic. Consider, a logic in which not all possible combinations of properties are allowed, as in a Boolean logic. There are many ways to restrict combinations; here we consider three very simple types of consistent pattern that an observer interested in distal causal structure might naturally focus on. The three types, taken together, suffice to define a reduced logic suitable for the extraction of aspects of causal structure. In particular the logic will take the form of an algebra, in which causally coherent models of the environment are expressible as algebraic expressions. Interpretation of an observed worlds such as $\mathcal{W}$ above is then accomplished by finding the minimal solution to a certain inequality in the algebra.

(1) One obvious type of distal causal structure consists of one abstract property causally entailing another; this would correspond in the observables to one feature logically implying another:

$$\sigma_i \rightarrow \sigma_j, \quad (4)$$

where $\rightarrow$ is some transitive, reflexive binary relation on $\Sigma \times \Sigma$. The observer's implicit underlying theory that the environment is "causally coherent" corresponds to a tangible belief that such implicational constraints have a tendency to exist and to be maintained consistently throughout a given world. Hence it is natural to express a given set of observations in terms of which pairs of properties (σ_i, σ_j) obey $\sigma_i \rightarrow \sigma_j$.

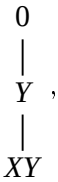
What patterns of property covariation does such implication impose on a property set? Consider a property set $\Sigma = \{X, Y\}$. Without any constraint, the possible objects in this world comprise the entire lattice of subsets of Σ:



(5)

in which each edge connects one object to another that contains exactly one more property, and 0 denotes the empty property string. This lattice, in effect, fully expresses the structures extant in a given world, and ranks them by "degree of structure." Lattices (and related partial orders) allow for a rich representation of such interpretations; see [12, 5, 6, 7, 27, 28, 14].

Now we introduce the implicational constraint $X \rightarrow Y$. This rules out the combination X, yielding the reduced lattice:



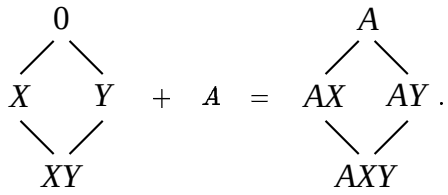
(6)

This lattice is a *sublattice* of the previous one, i.e. a restriction of the lattice partial order to a subset. A wide variety of other shapes of lattices can be obtained by adding other constraints in larger property sets, but, critically, not all patterns are possible. In particular only so-called "distributive" lattices occur (see [3] for an introduction, and [4, 9] for discussion of the significance of this restriction).

(2) A second type of consistent world structure occurs when one property σ holds consistently across all extant structures. Algebraically, this can be regarded as the result of "adding," or appending σ to each string in an observed set. (Recall that each object is really just a set of properties, so no duplicates are allowed.) For some set of objects $\{\mathcal{X}_i\} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots\}$, we write

$$\{\mathcal{X}_i\} + \sigma = \{\sigma\mathcal{X}_i | i = 1, 2, \ldots\}. \qquad (7)$$

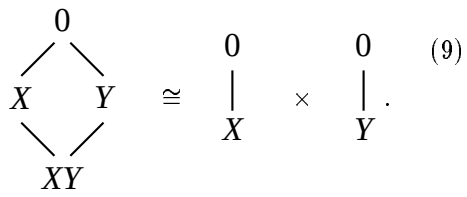
For example,



(8)

Addition of a fixed property does not change the shape of the resulting lattice, but simply adds a property to each node. Hence a distributive lattice plus a property is still a distributive lattice.

(3) The third and final component of the "structure algebra" is the necessary *a*causal component, namely orthogonal combination with causally irrelevant properties. Formally, such combination is simply the "direct product" (Cartesian multiplication of sets plus an induced partial order). Any fully Boolean lattice, such as Eq. 5 above, is isomorphic to the product of some number of two-element chains:



(9)

From our point of view, a Boolean lattice is simply a distributive lattice with no implicational constraints—i.e. a special case of distributive lattices. Moreover, the product of distributive lattices (including Boolean ones) is distributive.

Hence the three components of causal interaction reduce to the following "algebraic" form: take some intrinsic causal interaction $\omega \subseteq \Sigma \times \Sigma$, consisting of a set of implicational constraints on Σ; multiply the entailed distributive lattice by some set of orthogonal properties $\beta \subseteq \Sigma$; and then add some set of properties $\alpha \subseteq \Sigma$. The result is a distributive lattice of the form

$$\alpha + [\beta \times \omega]; \tag{10}$$

this expression captures the general form of a system that interacts causally in the way our hypothetical observer expects systems to behave.

Now, working backwards from an observed world $\mathcal{W}$, projected into a property set Σ, the observer would seek to find the (unique) minimal solution $\hat{\beta}, \hat{\alpha}, \hat{\omega}$ to the expression

$$\mathcal{W}|_{\Sigma} \subseteq \alpha + [\beta \times \omega], \tag{11}$$

which represents the default interpretation of the causal structure of the observed world $\mathcal{W}$. Given some extremely general formal assumptions [9] the solution to this inequality is the most *preferred* interpretation of the observed pattern, in that preference for any different solution consistent with the observations would be unstable or inconsistent.

The solution to Eq. 11 has three components: a component of *ever-present* properties $\hat{\alpha}$; a component of *causally irrelevant* properties, $\hat{\beta}$; and, most interestingly, a component of *causally interacting* properties $\hat{\omega}$. For a given set of observations, it is really $\hat{\omega}$ that captures the structure in the situation: an empty $\hat{\omega}$ indicates a completely structure-less environment in which all variables interact orthogonally. Hence in the sequel we will be primarily interested in the structure of the solution $\hat{\omega}$; when it is desirable to emphasize its origins, $\hat{\omega}$ obtained as part of the minimal solution to Eq. 11 on a set of feature strings $\{\mathcal{X}_i\}$ will be denoted by $\hat{\omega}(\{\mathcal{X}_i\})$. Notice that $\hat{\alpha}, \hat{\beta}$ and $\hat{\omega}$ are disjoint, and their union is Σ—that is, solving Eq. 11 really amounts to partitioning Σ into constant, causal, and acausal components.

Note that the Eq. 11 is not an equality, because as mentioned above not all observed worlds exhibit a coherent causal structure exactly. The inferred world, that is, can exhibit a more coherent structure than do the actual observations, yielding a "regularized" interpretation. This regularization and its formal conjugates may be regarded as accounting for the general tendency of human observers to simplify interpretations, even to the point of ignoring portions of the data—but somehow fixing on just the right generalization.

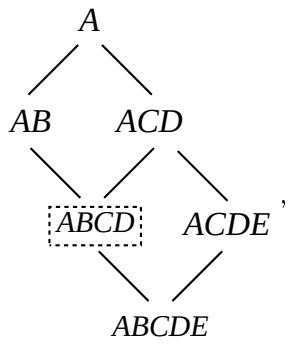
It is instructive to consider the resemblance between Eq. 11 and the equation for a straight line (or plane or hyperplane): in the analogy, the constant term α plays the role of the constant intercept, the implication set ω plays the role of slope, and the orthogonal term β extrudes the line out into additional dimensions. In this sense, the interpretation of structure entailed by Eq. 11 can be thought of as the observer regressing a set of observations to a linear (i.e., pure-implication) model.

Example 1: The amoeba world. Now

we return to the amoeba world of Fig. 1. The minimal solution here is

$$\begin{aligned} \hat{\alpha} &= A \\ \hat{\beta} &= B \\ \hat{\omega} &= E \to C, C \to D, D \to C. \end{aligned} \tag{12}$$

This solution corresponds to the lattice:



(13)

which is a sublattice of 2^{Σ}, equivalent to the expression

$$A \quad + \quad \left[\begin{array}{ccc} & 0 & \\ & | & \\ 0 & & 0 \\ | & & | \\ B & \times & CD \\ & & | \\ & & CDE \end{array} \right]. \tag{14}$$

In either form, the solution has a natural interpretation. The α component means that all objects in this world are `blob_like` (A). Notice that the regularity that all objects in this world are drawn from the same general type (blob-shaped objects), which is intuitively obvious to the human observer, is captured by the algebraic solution. This obvious aspect of the structure of the problem would not be captured by conventional procedures not oriented towards the recovery of world structure.

Continuing, the $\hat{\beta}$ component means that being `shaded` has no causal meaning—this property is not a cue to any other structure in this world.

The ω component is most intriguing. It contains a *cycle*, i.e. a set of σ_i's that imply each other:

$$\texttt{has\_nucleus} \leftrightarrow \texttt{has\_dotted\_membrane}. \tag{15}$$

This cluster of mutually correlated properties suggests a "mode," or, one might say, a *species*: a subpopulation in which certain properties consistently co-occur [25, 18]. These will be discussed in more detail below. Secondarily, some though not all of these dotted-membraned, nucleated objects are `large`; again this is not an arbitrary relationship, but rather the former is a precondition of the latter.

Notice that according to the inferred interpretation, one node in the above lattice, indicated by a box, is actually *missing* from the observations. That category, the observer infers, is possible under the underlying causal structure of this world, and ought to occur occasionally. This is a direct manifestation of how the the weakness of the logic enables the observer to make strong inductive inferences about the world, and how these inferences are regularized with respect to observations.

Example 2: A "Bongard" problem. A second example of inference using the "structure algebra" is provided by the classic visual induction problems of Bongard [2], often held up as a benchmark of intelligent generalization. Each problem consists of twelve panels such as those shown in Fig. 3, six on the left and six on the right. The problem posed to the observer is to generalize from the examples given to the presumably infinite classes

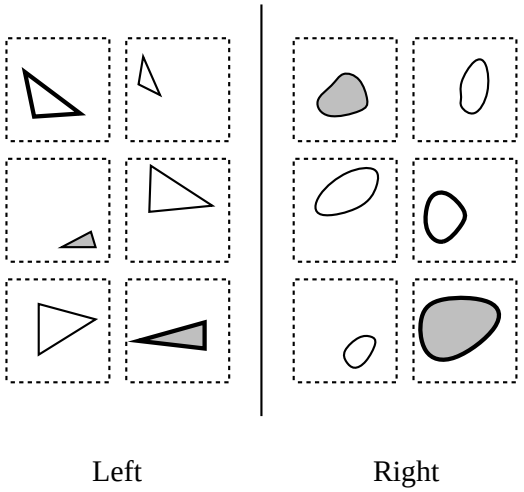


Figure 3: A typical "Bongard problem." What is the difference between the left and the right?

exemplified by the two groups, somehow ignoring the concrete but nevertheless uninteresting distinctions among panels on each side (e.g. some are larger, some smaller, some shaded, some not, etc.).

The dichotomy between the two groups is conveniently labeled ("left" and "right") by an oracle, thus making this an example of a supervised induction problem. The essential problem here is to fix on the correct properties suitable for distinguishing the two groups; we will not address this problem here. Rather, we focus on how the algebra expresses the structure inherent in the Bongard set-up.

The canonic Bongard problem is in essence constructed as follows: twelve objects of some general type A, of which the six on the left have some property X, while the six on the right have $\neg X$. On top of this, some distractor properties B are added, which interact orthogonally with X on both left and right sides.

Of course this composition is reminiscent of the composition of the structure algebra itself, and it is no surprise that the algebra captures it neatly. To see how, consider that the oracle's labels "left" and "right" can be regarded as another feature, coded as L and $\neg L$. Now the Bongard problem can be regarded as a world $\mathcal{W}_B$ in exactly the sense defined above. The structure of this world is captured when it is projected into the alphabet $\Sigma_B = \{A, B, L, X\}$, in that

$$\hat{\omega}(\mathcal{W}_B|_{\{A,B,L,X\}}) \supseteq \{L \to X, X \to L\}. \quad (16)$$

The cycle $L \leftrightarrow X$, contained in the solution $\hat{\omega}$, captures the fact that in the Bongard world the critical property X consistently occurs on the left, and consistently fails to occur on the right. Moreover, this statement holds for any larger alphabet $\Sigma'_B \supseteq \Sigma_B$. The structure inherent in the Bongard world is captured cleanly in ω-space.

Example 3: An unlabeled Bongard problem. The above analysis of the labeled Bongard problem reduces the oracle's labels "left" and "right" to a property in Σ. Hence it makes sense to regard an *un*labeled problem isomorphically, just so long as there is at least one property playing the role of the label L in the labeled case. All that is required is that this property exhibit a consistent correlation (mutual implication) with the critical property X. The structure inherent in the world plays the role of oracle.

Consider the new population of amoeba-like objects shown in Fig. 4. It is intuitively clear that there are two categories of object here: one whose members are consistently C, D, and E (i.e. `has_nucleus`, `has_dotted_membrane`, and `large`), and other whose members are consistently not. Yet this simple and obvious inference is not easily captured by existing categorization algorithms, which generally either have to be fed a number of categories, or split data into new categories only on the basis of

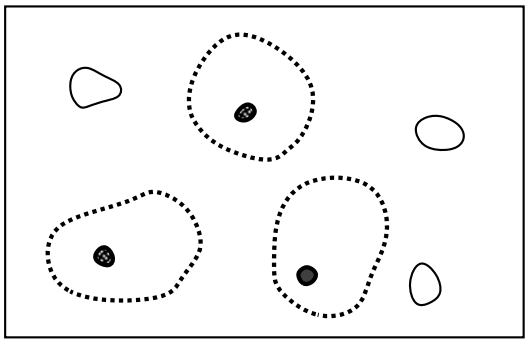


Figure 4: An alternative population of "amoebae," amounting to an unlabeled Bongard problem.

ad hoc rules. Algebraically, though, the solution here is

$$\begin{aligned} \hat{\alpha} &= A \\ \hat{\beta} &= \emptyset \\ \hat{\omega} &= C \to D, D \to C, \\ &\quad\ C \to E, E \to C, \\ &\quad\ D \to E, E \to D. \end{aligned} \tag{17}$$

This is a very structured solution—i.e. it has a large $\hat{\omega}$. In particular it has a 3-property cycle that clearly corresponds to a mode or subspecies, and points to a systematic distinction, of exactly the same type as the left-right distinction in the labeled Bongard problem. The inherent structure in the problem, rather than explicit labels provided by an oracle, is providing the necessary inferential leverage to recover the categories. In this sense, the structure algebra, and in particular the structure to be discovered in ω-space, dissolves the boundary between supervised and unsupervised learning; the intelligent observer takes advantage of the structure inherent in the problem wherever it is to be found.

3 Modal forms in inferred structure

We now shift out attention from structure in the world to structure in the induced ω-space, that is, from the structure of $\mathcal{W}$ to the structure of $\hat{\omega}(\mathcal{W}_\Sigma)$. In particular, we are interested in modal forms in $\hat{\omega}$: structures whose existence human observers seem to presuppose. Such structures, it is argued, constitute the inductive bias that accounts for human observers' adept interpretation of complex systems.

We focus on two patterns: *cycles* and *disjunctions.*

Cycles. Among those who study human concepts, it is now a common (though not entirely uncontroversial) view that mental models of categories in the world (and very possibly the categories themselves) depend critically on property correlation—the tendency of some properties to occur with greater consistency in the presence of certain other properties (see [25, 26, 18]). Human interpretations of such correlations are prone to bias: we tend to seek positive examples of property correlations that we believe hold, while ignoring actual correlations that are observed but do not fit into any existing mental framework [15]. The bias towards such structure is natural for human observers, existing as they do in a world populated by well-defined, highly redundant subpopulations ("prototypes") of objects exhibiting arbitrarily large sets of consistent features (the "Principle of Natural Modes" [1, 22]).

In the framework of the structure algebra, such prototypes naturally take the form of cycles in ω: subsets of Σ all of whose members imply each other under $\to$. As mentioned above, any such cycle can be thought of as a "species:" a subpopulation exhibiting consistent structure. The crucial insight is that in a

natural environment, one can expect such cycles to be *arbitrarily* large: a genuinely distinct subtype can be expected generally to be distinct from other subtypes whenever new properties are considered.

Disjunctions. When multiple such subtypes exist within a given world $\mathcal{W}$, they correspond to multiple *disjoint* (non-overlapping) clusters in ω; this is the generic case. (These clusters are not necessarily cycles, because implication is one-way, but each may have a single cycle at its center.) Abstracting, we focus on these disjoint sets and define a notation to summarize them.

For a fixed property set Σ and ω, denote by

$$\mathcal{C}(\omega) = c_1, c_2, \ldots \tag{18}$$

the partition of Σ into sets that are disjoint in ω: i.e. $\sigma_1 \in c_1, \sigma_2 \in c_2$ implies $\sigma_1 \not\rightarrow \sigma_2$ and $\sigma_2 \not\rightarrow \sigma_1$. Given an observed world $\mathcal{W}$, $\mathcal{C}(\hat{\omega})$ provides a list of the observable subtypes in $\mathcal{W}$. Hence we call $\mathcal{C}(\hat{\omega})$ the *typology* of $\mathcal{W}$ under Σ, and each c_i a *type*.

A sufficient condition for the recovery of essential structure. We now are in a position to ask concretely an essential question posed more vaguely earlier in this paper: under what circumstances does a feature set allow the "essential" structure of a world $\mathcal{W}$ to be recovered?

First, we define a mapping that expresses how the structure recovered from one property set changes when the same world is projected into another property set. For two property sets $\Sigma_1 \supseteq \Sigma_2$, and a typology $\mathcal{C}_1$ defined on Σ_1, define the mapping

$$\pi : \mathcal{C}_1 \mapsto \mathcal{C}_2 \tag{19}$$

as the *restriction* of $\mathcal{C}_1$ to the smaller property set Σ_2; this is the natural projection of $\mathcal{C}_1$ into $\mathcal{C}_2$. This projection maps one typology into another which is less informative, in that it is expressed in the more impoverished language Σ_2. Given the discussion above of the cycles as subtypes within an extant population, though, one might expect that the projection might sometimes preserve *all* that is essential about the larger, richer language Σ_1. This happens when the following condition is satisfied:

(Projection Condition) Given $\Sigma_1 \supseteq \Sigma_2$, and a fixed world $\mathcal{W}$, the projection

$$\pi : \mathcal{C}[\hat{\omega}(\mathcal{W}|_{\Sigma_1})] \mapsto \mathcal{C}[\hat{\omega}(\mathcal{W}|_{\Sigma_2})] \tag{20}$$

is an isomorphism.

This admittedly strong condition captures what we mean by Σ_1 and Σ_2 revealing the "same" structure in the world. The types extracted by the structure algebra via Σ_1 are, one for one, the *same* types as are extracted via Σ_2. Each type (cluster) may be reduced in size, but not so much that it has disappeared completely.

One final speculation completes this superficial investigation of the structure of ω-space. Statisticians define the "true" value of a parameter by means of a convenient abstraction, the value the parameter in the infinite population—i.e., its value in a finite sample as the sample is allowed to get larger without limit. Analogously, we might define the "true" structure of ω for a given world $\mathcal{W}$ as the structure of $\hat{\omega}(\mathcal{W}|_{\Sigma})$ as Σ is allowed to get larger without limit—that is, as the expressive power of the description language grows without bound.

Consider the typology $\mathcal{C}_0$ of this idealized, "true" ω. Now consider a property set Σ that satisfies the projection condition with respect to $\mathcal{C}_0$; that is, Σ such that

$$\pi : \mathcal{C}_0 \mapsto \mathcal{C}[\hat{\omega}(\mathcal{W}|_{\Sigma})] \tag{21}$$

is an isomorphism. Such a Σ is *complete* with respect to the world $\mathcal{W}$: it allows all of the $\mathcal{W}$'s structure to be revealed to the observer. This notion of completeness is comparable to the idea of "identification in the limit" of Gold [10] (see also [16]). Completeness in this regard does not mean that the description language is universal; rather, only that it is sufficiently rich to support a certain kind of structure-preserving mapping between the world and the observer's inferences.

In this view, human observers' great success in comprehending the structure of the natural world can be attributed to our possession of the right Σ—a description language that is complete, in this technical sense, with respect to the true algebraic structure of the particular world in which we evolved. For automatic interpretation systems to achieve the same kind of facility with artifactual worlds—or indeed with any closed worlds—it would suffice for them to be endowed with similarly complete description languages. Finding such languages remains a daunting problem; the formalism presented here allows a precise characterization of what it would mean for this problem to be solved.

4 Conclusion

Researchers in fields concerned with automatic generalization and induction have, historically, tended to emphasize domain-independent procedures, regarding reliance on domain-specific knowledge as a circumvention of the inherent induction problem. But this may well be a false dichotomy. The approach taken here is neither domain-specific nor domain-independent; rather, we attempt to characterize the canonic forms that domain-*specific* structure tends to take *across* domains. Different domains—in the terms used here, different worlds—share the tendency to exhibit consistent causal structure. This paper has taken a step towards characterizing this causal structure in a general way, while articulating minimal conditions for its recovery by an observer. Intriguingly, these conditions are not expressed as constraints on the world $\mathcal{W}$, *nor* on the language Σ used to describe that world; but rather on the *relationship* between $\mathcal{W}$ and Σ.

Clearly, the causal structure of worlds can be substantially more complicated than the minimal case of pairwise implication investigated here. Intriguingly, though, this simple model can account for a wide variety of hitherto mysterious human intuitions, particularly in perceptual domains, where the goal is to achieve a theory of "perceptual semantics." Perceptual grouping, like the problem discussed above, is classic case of a problem in which human observers succeed effortlessly and easily out-perform conventional algorithmic solutions to an analogous problem (statistical clustering). Formal machinery closely related to that presented in the current paper (really a hierarchical, multi-resolutional generalization of that presented here) accounts for human interpretations with remarkable success [9, 8]. Efforts to extend this success to a wider range of perceptual problems are underway (see [9, 24] for pointers to future directions). This research is particularly exciting for the prospect it holds in forging a tighter mathematical link between models of perceptual interpretation and models of higher-level cognitive processes; and between models of human inference processes and models of automatic inference.

References

[1] A.F. Bobick. Natural object categorization. Technical Report 1001, Massachusetts Institute of Technology Arti-

ficial Intelligence Laboratory, November 1987.

[2] M. Bongard. *Pattern recognition.* Spartan Book, New York, 1970.

[3] B.A. Davey and H.A. Priestley. *Introduction to lattices and order.* Cambridge University Press, Cambrdige, 1990.

[4] M. Erné. Distributive laws for concept lattices. *Algebra Universalis*, 30:538–580, 1993.

[5] J. Feldman. Perceptual simplicity and modes of structural generation. In *Proceedings of the 13th Annual Conference of the Cognitive Science Society*, 1991.

[6] J. Feldman. Constructing perceptual categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1992.

[7] J. Feldman. *Perceptual categories and world regularities.* PhD thesis, Massachussetts Institute of Technology, 1992. Available as Technical Report #6, Rutgers Center for Cognitive Science.

[8] J. Feldman. Perceptual models of small dot clusters. In *Proceedings of the DIMACS Workshop on Partitioning Data Sets.* American Mathematical Society, 1993.

[9] J. Feldman. Regularity-based perceptual grouping. Technical report, Rutgers Center for Cognitive Science, 1995.

[10] E. Gold. Language learning in the limit. *Information and control*, 10:447–474, 1967.

[11] A. Jepson and W. Richards. What is a percept? Occasional Paper 43, MIT Center for Cognitive Science, April 1991.

[12] A. Jepson and W. Richards. A lattice framework for integrating vision modules. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(5):1087–1096, September/October 1992.

[13] M. Leyton. *Symmetry, Causality, Mind.* M.I.T. Press, Cambridge, MA, 1992.

[14] N. Moray. A lattice theory approach to the structure of mental models. *Phil. Trans. R. Soc. Lond. B*, 327:577–583, 1990.

[15] G.L. Murphy and E.J. Wisniewski. Feature correlations in conceptual representations. In G. Tiberghien, editor, *Advances in Cognitive Science*, volume Vol. 2: Theory and Applications, pages 23–45. Chichester Ellis Horwood, 1989.

[16] D. Osherson, M. Stob, and S. Weinstein. *Systems that learn.* M.I.T. Press, Cambridge, MA, 1986.

[17] J. Pearl and T.S. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 1–12. Morgan Kaufmann, San Mateo, CA, 1991.

[18] W.V.O. Quine. Natural kinds. In H. Kornblith, editor, *Naturalizing Epistemology.* M.I.T. Press, Cambridge, MA, 1985.

[19] R. Reiter. On closed world data bases. In H. Gallaire and J. Minker, editors, *Logic and data bases.* Plenum Press, New York, 1978.

[20] R. Reiter and A. K. Mackworth. A logical framework for depiction and image interpretation. *Artificial Intelligence*, 41:125–155, 1989.

[21] W. Richards. *Natural computation.* M.I.T. Press, Cambridge, MA, 1988.

[22] W. Richards and A. Bobick. Playing twenty questions with nature. In Z.W. Pylyshyn, editor, *Computational Processes in Human Vision: An Interdisciplinary Perspective*, pages 3–26. Ablex Publishing Corporation, Norwood, NJ, 1988.

[23] W. Richards and A. Jepson. What makes a good feature? In L. Harris and M. Jenkin, editors, *Spatial Vision in Humans and Robots.* Cambridge University Press, 1992.

[24] W. Richards, A. Jepson, and J. Feldman. Priors, preferences, and categorical percepts. In D. Knill and W. Richards, editors, *Perception as Bayesian Inference.* Cambridge University Press, Cambridge, U.K., 1995.

[25] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[26] E. Smith and D. Medin. *Categories and Concepts.* Harvard University Press, Cambridge, MA, 1981.

[27] S. Watanabe. *Knowing and guessing: a quantitative study of inference and information.* John Wiley, New York, 1969.

[28] S. Watanabe. *Pattern recognition: human and mechanical.* John Wiley, New York, 1985.

Acknowledgments

I am grateful to Whitman Richards, Allan Jepson, and Alex Meystel for useful comments and discussions. The research reported in this paper was made possible by the Rutgers Center for Cognitive Science.