

Inference and coherence in causal-based artifact categorization

Guillermo Puebla

Universidad de Tarapacá
General Velásquez 1775
Arica, Chile

&

Sergio E. Chaigneau

Centro de Investigación de la Cognición
Facultad de Psicología
Universidad Adolfo Ibáñez
Diagonal Las Torres 2640, Peñalolén
Santiago, Chile

May, 2013

Running head: artifact categorization

Corresponding author:

Sergio E. Chaigneau
Escuela de Psicología
Universidad Adolfo Ibáñez
Diagonal Las Torres 2640
Peñalolén, Santiago, Chile
Phone: +562 2331 1199
Email: sergio.chaigneau@uai.cl

Abstract

In four experiments, we tested conditions under which artifact concepts support inference and coherence in causal categorization. In all four experiments, participants categorized scenarios in which we systematically varied information about artifacts' associated design history, physical structure, user intention, user action and functional outcome, and where each property could be specified as intact, compromised or not observed. Consistently across experiments, when participants received complete information (i.e., when all properties were observed), they categorized based on individual properties and did not show evidence of using coherence to categorize. In contrast, when the state of some property was not observed, participants gave evidence of using available information to infer the state of the unobserved property, which increased the value of the available information for categorization. Our data offers answers to longstanding questions regarding artifact categorization, such as whether there are underlying causal models for artifacts, which properties are part of them, whether design history is an artifact's causal essence, and whether physical appearance or functional outcome is the most central artifact property.

Keywords: causal-based categorization; artifacts; essentialism; coherence effect; causal inference

1. Introduction

One theoretical goal in psychology has been to describe how young children and adults categorize artifacts (for reviews, see German, Truxaw, & Defeyter, 2007; Hernik & Csibra, 2009; Kelemen & Carey, 2007; Malt & Sloman, 2007; Oakes & Madoles, 2008). The current work was motivated by existing theoretical and empirical differences between natural and artifact kinds regarding modes of causal categorization. Frequently, these kinds are set apart based on the way their respective causal models are configured (Ahn, Kalish, Gelman et al., 2001; Ahn & Kim, 2001). For example, several researchers have proposed that natural kinds conform to a common cause model, in which internal properties like DNA cause many other observable properties such as appearance and behavior (Gelman, 2003). For artifact kinds, some researchers have also proposed a common cause causal model, in which design history is the root cause of many others artifact properties like its physical structure and function (Ahn, 1998; Ahn, Kalish, Gelman et al., 2001; Kemler-Nelson, Frankenfield, Morris, & Blair, 2000), while other researchers have proposed a common effect model, in which many properties such as physical structure and agent actions contribute to the artifact achieving its function (Chaigneau, Barsalou, & Sloman, 2004).

Given the theoretical relevance of causal knowledge for artifact kinds, it seems natural to examine artifact categorization as a case of causal-based categorization. Causal categorization theories hold that causal relations that are thought to be true for a category affect category membership judgments (Ahn, Kim, Lassaline & Dennis, 2000; Sloman, Love, & Ahn, 1998; Rehder, 2003a, b). Two key mechanisms by which causal knowledge affects categorization are coherence and inference. When categorizing by coherence (Rehder & Kim, 2006), people estimate if a configuration of known properties (i.e., an exemplar) could be generated by the category's causal model, and then use that estimate to categorize. Imagine the simplest possible causal model. Imagine that the bird category's causal model reduces to knowledge that having large wings

causes the animal to be able to fly ($A \rightarrow B$). When this causal link is conceptualized as deterministic, coherence predicts that A and B will be similarly relevant for categorization, and importantly, that A and B will interact (slightly different predictions would be derived if the link were probabilistic). An interaction means that because $A \rightarrow B$ implies a correlation, exemplars that violate the expected correlation (i.e., a bird with large wings that does not fly, or a bird with small wings that does) will be categorized as poorer category members than exemplars that preserve the expected correlation (i.e., a bird with large wings that flies, or a bird with small wings that does not). Note that a categorizer that considered A and B as isolated properties would be insensitive to correlations, with A and B contributing independently to categorization. When categorizing inferentially (Rehder & Kim, 2009), people use known properties to infer the presence of other causally related properties that are themselves important for categorization. These inferences can be diagnostic (i.e., using a known effect to infer the presence of its cause; which we will call here retrospective inferences, to emphasize the inference's direction) and predictive (i.e., using a known cause to infer the presence of its effect; which we will call here prospective inferences) (see Fernbach, Darlow, & Sloman, 2011). Note that these inferences are important when categorizing exemplars where category relevant properties are unknown.

Regarding the issue of differences in causal categorization between artifacts and natural kinds, Hampton, Storms, Simmons and Heussen (2009) have proposed a more radical difference than those suggest by other researchers. In two experiments, Hampton et al. showed that when people categorized biological kinds, properties interacted when combined (i.e., categorized by coherence, as discussed above), but that when people categorized artifacts no such interactions obtained, and properties contributed independently to exemplar categorization instead (inconsistently with the coherence mechanism). To these authors, their results suggested that

artifact kinds lack strong coherent underlying causal models, which is why their participants did not naturally resort to coherence when categorizing them.

However, it's not completely clear to us what the implications of Hampton et al.'s (2009) results are. Not using coherence to categorize does not imply that causal knowledge is not available for categorization. Categorizing by retrospective and prospective inferences also requires using causal knowledge to guide those inferences (Rehder, 2010). In this context, Hampton et al.'s results do not automatically imply an absence of an underlying causal model. It may be that artifacts are categorized by inferences about the presence of certain relevant but unobserved properties; inferences that are guided by causal knowledge. Accordingly, our main goal in the current experiments was to test whether participants categorized artifacts by coherence or by inferences (or perhaps both), and to understand the causal relations, if any, that our participants used to guide their categorization judgments.

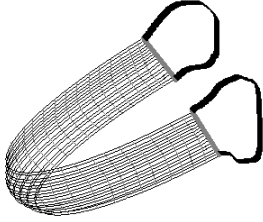
2. Experiments' overview

In the current experiments, participants were presented with scenarios describing a novel artifact's design history (H), its physical structure (P), an agent's goal when using the artifact (G), an agent's action (A), and the functional outcome (O). These variables are close to exhausting properties that are considered relevant for artifact categorization across many studies. After receiving information about the state of these properties, participants were asked to make a category membership rating. Because properties could be intact, compromised or absent, we could construct multiple scenarios that combined these property values in different patterns. On all four experiments, ratings were analyzed using the regression method (Rehder & Hastie, 2001). In this method, participants provide category membership ratings for all possible property combinations, allowing the computation of individualized regression equations. The basic

procedure in our experiments required presenting participants with combinations of binary variables corresponding to intact and compromised versions of H, P, G, A and O: H could describe the artifact being designed towards functionality x or functionality y ; P could be described as adequate to achieve functionality x , or not adequate to achieve it; G could be described as intentional and coherent with functionality x , or accidental and not coherent with functionality x ; A could be described as coherent with functionality x , or not coherent with functionality x ; O could be described as achieving or not functionality x (for an example, see Figure 1; for different causal relations that could be established among H, P, G, A and O, see Figure 2).

In the regression method, if participants were to receive scenarios with n binary properties, they would receive and rate 2^n scenarios. When intact and compromised properties are coded, respectively as 1 and -1, these values can be entered into individualized regression equations to predict a participant's categorization ratings. Furthermore, 2-way and higher order interaction terms can be computed by entering the product of the corresponding property values as predictors into the equations. Thus, in the simplest hypothetical scenario with only two binary valued properties, the interaction term for scenarios that showed both properties intact would be $1 \cdot 1 = 1$, for those showing one property intact and one compromised would be $1 \cdot -1 = -1$, and for those showing both properties compromised would be $-1 \cdot -1 = 1$. If this interaction term came to be positively correlated with categorization ratings, it would mean that participants viewed exemplars with coherent properties (i.e., both properties in the same state) as better category members than exemplars with incoherent properties. Because individualized regression equations yield coefficients for single properties and interaction terms, these regression coefficients can then be used as individual data points reflecting, across participants, the contribution of each predictor variable to the ratings. The distribution of coefficients across participants can then be submitted to significance tests.

In an ancient culture, a settler called Kne-Mû wanted to make an object to catch small fish living in large numbers in certain streams. Because he didn't have an object to do that, he decided to make it. The object consisted of a series of intertwined vegetable fibers. On each side, the object had handles (as shown in the drawing).



One day, another settler called Knat-knê wanted to catch some small fish from a stream. He found the object Kne-Mû made and thought that it would be useful for catching fish. Knat-knê grasped the object by both handles and kept it stretched just below the stream's surface. As a result of the events described, fish were trapped in the artifact.

Question: Would you say that this object is a fish-catcher?

Figure 1. Baseline scenario for the fish-catcher artifact in Experiment 1.

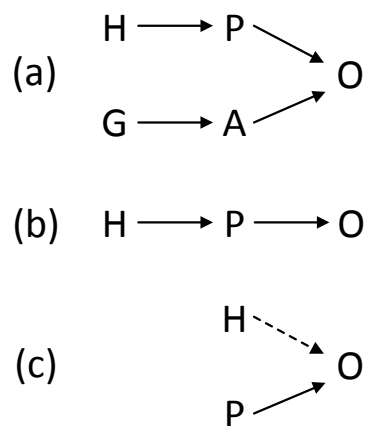


Figure 2. Panel (a) shows a possible causal model that includes all properties used across the current experiments. Panels (b) and (c) show two possible causal models used by our participants, based on results for Experiment 1. The dotted line in panel (c) reflects a weak causal link from H to O.

Another possibility that the regression method affords is to detect whether a given property allows inferences about the state of properties for which information is lacking. Imagine again the simplest hypothetical scenario with only two properties (e.g., P and O), where

sometimes the scenario is presented without information about property O (i.e., only information about P is provided). If P's correlation with categorization ratings derives from it being used to make inferences about the state of O, with the latter being the important property for categorization, then P's correlation with categorization ratings should increase when no information about O is provided. In other terms, because P is used to infer the state of O, which is then used for categorization, when information about O is absent, P provides all the information used for categorization. To perform this analysis, scenarios should vary the state of P (intact and compromised) when information about O is available and when information about O is absent.

In all our experiments, baseline scenarios (i.e., all components in accordance with functionality x) were always rated first to provide an anchor for participants' ratings. As will be explained later, in all experiments we also performed model fitting to provide additional evidence of participants' categorization process.

3. Experiment 1

The main issue we wanted to address in this experiment was whether participants categorized artifacts depending on coherence (i.e., considered better exemplars those with coherent property states) or depending on the state of individual properties. In this sense, this experiment is a conceptual replication of Hampton et al. (2009). Computing and testing weights for interaction terms allowed us to answer this question.

A second and related issue was the following. In previous experiments using stimuli similar to those used here (Chaigneau, Barsalou, & Sloman, 2004) it was assumed that people used available information about an artifact's history and current use to infer the expected outcome (O, which was not provided), and then used that inferred outcome to answer questions about function, causality and categorization. However, the condition where the outcome was known

was not tested in those experiments. The current experiment is precisely that condition. If the process model in Chaigneau, Barsalou and Sloman (2004) is correct, then when information about O is provided, O should overwhelm other properties regarding its weight for categorization. Thus, in this experiment participants were presented with scenarios containing information about H, P, G, A and O, where each could adopt one of two values (i.e., intact or compromised).

3.1. Method

3.1.1. Design and Participants

This experiment used a 3 (novel artifacts) x 32 (scenarios) mixed design. Thirty Adolfo Ibáñez and Tarapacá University undergraduates participated in this study (native Spanish-speakers).

Participants signed informed consent and were randomly assigned to one of 3 artifacts and one of 5 random orders of scenarios.

3.1.2. Materials

Three novel artifact categories were tested (“peinador”, “cazador de peces” and “tatuador”; respectively, “hair-brusher”, “fish-catcher” and “tattoo-maker”), with 32 scenarios for each category. Each category was designed to afford two plausible functions, one serving as cue to name the artifact. For example, the fish-catcher consisted of a net of vegetable fibers which could (in principle) be used either to catch fish or to carry stones. The cue function was fixed across all scenarios so the question was always the same (e.g., would you say that this object is a fish-catcher?). Scenarios described one character that created an object and a second character that used it. A graphic depiction of the artifact’s physical structure was included in all scenarios. As an example, Figure 1 shows the fish-catcher scenario specifying all elements as adequate (i.e., baseline). When H was compromised, the designer created the object for the non-cue function, but the second character used it for the cue function (e.g., a net designed to carry stones which is

then used as a fish-catcher). When P was compromised, the artifact's physical structure was described and depicted as not affording its cue function (e.g. a net with several holes on it). When G was compromised, the second character's actions were described as accidental (e.g., the second character performed the appropriate actions but was acting distracted and not intending to catch fish). When A was compromised, the second character was described as not performing the appropriate actions (e.g., shaking the net just under the water's surface instead of keeping it stretched and still). When O was compromised, the functional outcome associated with the cue function was not achieved (e.g., fish were not trapped). Thus, the 32 scenarios for each category presented participants with all combinations of adequate and compromised H, G, A, P and O ($2^5 = 32$).

3.1.3. Procedure

Stimuli were presented on a computer screen at a self-paced rate. Participants read instructions on the screen, but also heard them read aloud by the experimenter. Participants received two training scenarios, which described the creation and use of a hammer. One of these scenarios was a baseline (i.e., all properties adequate), and the second scenario presented the opposite extreme of the scale (i.e., all components compromised). The goal of this training was to teach participants to use the breadth of the scale. Ratings were performed on a 7-point scale, with 1 always reflecting the low-end ("no") and 7 the high-end of the scale ("yes").

3.2. Results

3.2.1. Results for the regression analysis

In this and the next experiments, individual ratings were logarithmically transformed prior to the analyses. There is much evidence that properties are multiplicatively, rather than additively, integrated in categorization. Several categorization models successfully use multiplicative

integration to account for data (e.g., Kruschke, 1992; Medin & Schaffer, 1978; Minda & Smith, 2002; Nosofsky, 1984, 1986), which means that increasing the number of cues relates non-linearly to measures such as categorization ratings. Note that this non-linearity may be difficult to distinguish from a coherence effect. To prevent this problem, logarithms transform multiplicative relations into additive ones, so that property interactions may be more confidently attributed to coherence.

Properties were analyzed by performing a multiple regression for each participant. Five predictor variables (H, P, G, A and O) were coded as -1 if the feature was compromised and +1 if it was intact. The regression weight associated with each predictor represents the influence that each element had on ratings. We also computed 26 interaction terms and introduced them in the individualized regression equations (10 two-way, 10 three-way, 5 four-way, and 1 five-way interactions). Averaged regression weights over participants for individual predictor and two-way interaction terms are presented in Figure 3.

According to one-sample *t* tests, only the regression weights of H, P, O and PO were significantly different from zero ($t(29) = 5.64, p < .001$; $t(29) = 3.43, p < .001$; $t(29) = 7.7, p < .001$; $t(29) = -2.65, p < .05$, respectively). No other single predictor or interaction term was significantly different from zero (all $ps > .05$).

Individual regression weights were entered into a mixed 3 (artifact) x 5 (scenario order) x 5 (property: H, P, G, A, O) ANOVA with repeated measures on the last factor, which showed no effects of artifact or of scenario orders. Thus, we collapsed these factors for further analyses. A violation of the sphericity assumption was handled by correcting degrees of freedom with Huynh-Feldt's epsilon ($\epsilon = .634$). For clarity of presentation, degrees of freedom are presented here without adjustment. There was a main effect of property ($F(4, 116) = 21.90, MSe = .05, p < .001, R^2$

= .43, power > .99). Post hoc tests (with Bonferroni adjustment) showed that the regression weight associated with O was higher and significantly different from G, A and P (all p s < .05) but not significantly different from H (p > .05). Additionally, H was higher and significantly different from G and A (both p s < .05), but not from P (p > .05). There were no other significant differences.

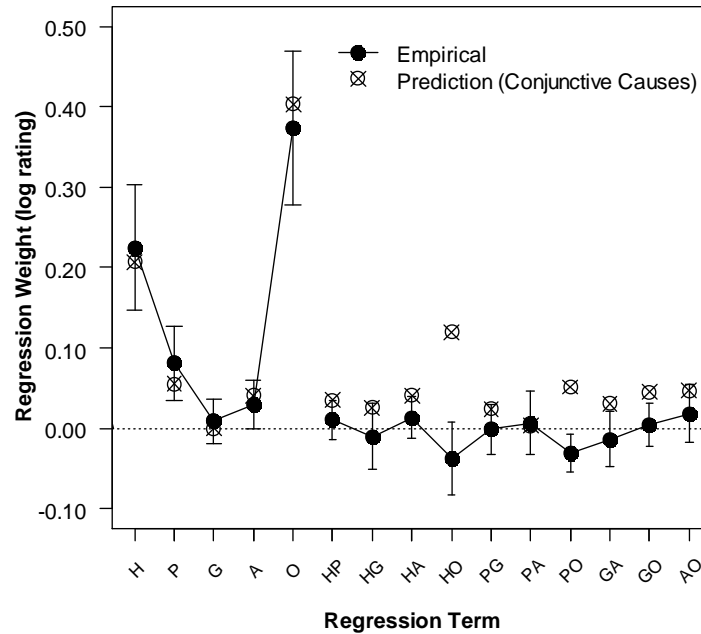


Figure 3. In Experiment 1, mean regression weights for history (H), agent goal (G), agent action (A), physical structure (P), outcome (O) and for two-way interaction terms. Other interactions are not shown due to space limitations. Only H, P, O and PO were significantly different from zero. Bars are 95% confidence intervals. Predictions derived from model fitting are shown superimposed on data (see text for further detail).

3.2.2. Model description

To further evaluate if participants categorized based on coherence, we fit the Generative Model (GM) of categorization (Rehder, 2003a, b; Rehder & Kim, 2006) to their responses. According to this model, causal relations among category features are represented as probabilistic causal mechanisms. When faced with a possible category member, people estimate how probably it could have been generated by the category's causal model in order to categorize. In the GM representation, a category k establishes a set of causal mechanisms. Each mechanism relates a

feature j with its parent i operating with probability m_{ij} when i is present. The model assumes that other background causes of j operate collectively with probability b_j . In the current studies, we use two alternative formulations according to the probabilistic relation among j 's parents. When j 's parents operate independently, then j 's parents and the background causes produce j in members of category k conditional on the state of j 's parents with probability:

$$p_k(j|\text{parents}(j)) = 1 - (1 - b_j) \prod_{i \in \text{parents}(j)} (1 - m_{ij})^{\text{ind}(i)} \quad (1)$$

where $\text{ind}(i)$ is an indicator variable that evaluates to 1 when i is present and 0 otherwise. The probability of a root cause r is a free parameter c_r .

In the current studies we also consider the possibility that j 's parents act as conjunctive causes (Rehder, 2011). In the case of conjunctive causation, several causes need to be present to bring about one effect. The GM represents this situation by relating j with its parents through a single causal mechanism. For example, consider a causal model in which h and i are conjunctive causes of j . In this case j is produced in members of category k conditional on the state of j 's parents with probability:

$$p_k(j|\text{parents}(j)) = 1 - (1 - b_j)(1 - m_{hij})^{\text{ind}(hi)} \quad (2)$$

where m_{hij} represents the probability that the causal mechanism relating parents h and i with j operates, and $\text{ind}(hi)$ is an indicator variable that evaluates to 1 when h and i are both present and 0 otherwise.

For the purposes of this analysis, we used two causal models (see Figure 4) that were intended to capture most of the possible causal models that could account for our results (i.e., by setting to zero a subset of causal links in these models, any model in Figure 2 could be retrieved).

In the *independent causes model* (Figure 4, panel A), H was treated as a deep cause of all other features in the model, G was a cause of P, A and O, and P and A were independent causes of O. The *conjunctive causes model* (Figure 4, panel B), was similar to the independent causes model, but instead of P and A being independent causes of O, they were treated as conjunctive causes.

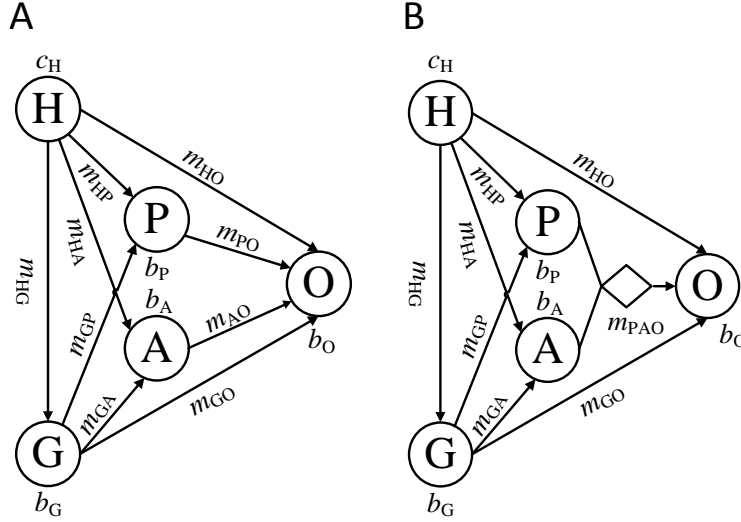


Figure 4. Independent Causes Model (panel A). Conjunctive Causes Model (panel B).

3.2.3. Model fitting

We fit both the independent and conjunctive causes models to our participants' log transformed responses. Specifically, participants' ratings were predicted according to the formula:

$$rating(E_i) = \beta p_k(E_i; c_i, m_{ij}, m_{hij}, b_i) \quad (3)$$

where E_i is a given exemplar and M is the category causal model (plus its parameters). Parameter β maps the model's predictions onto the rating scale according to a linear transformation.

Expressions for p_k were built applying iteratively equation (1) for the independent causes model and equations (1) and (2) for the conjunctive causes model. Parameters c_i , m_{ij} , m_{hij} and b_j were constrained to the range $[0,1]$. Parameter β was constrained to the range $[0,\infty]$.

Table 1

Parameter estimates and measures of fit for the conjunctive and the independent causes models. Note that causal parameters (m_{ij}) are all close to zero, but that base rate parameters for H and O (in bold) are the only properties clearly above chance (0.5). Parameter values left blank indicate parameters that are not in the corresponding model.

Parameters and measures	Model			
	Conjunctive Causes		Independent Causes	
	M	SE	M	SE
c_H	0.596	0.024	0.599	0.024
b_G	0.485	0.012	0.483	0.012
b_P	0.501	0.022	0.499	0.022
b_A	0.486	0.009	0.481	0.011
b_O	0.626	0.028	0.601	0.030
m_{HG}	0.047	0.012	0.052	0.013
m_{HP}	0.047	0.019	0.049	0.019
m_{HA}	0.062	0.019	0.050	0.016
m_{HO}	0.114	0.049	0.120	0.051
m_{GP}	0.045	0.017	0.044	0.017
m_{GA}	0.056	0.026	0.071	0.029
m_{GO}	0.119	0.039	0.175	0.055
m_{PO}	-	-	0.124	0.053
m_{AO}	-	-	0.113	0.044
m_{PAO}	0.158	0.060	-	-
β	34.598	1.807	34.495	1.823
Average <i>SSE</i>	6.303		6.069	
Average <i>RMSE</i>	0.541		0.548	
R^2	0.810		0.795	

Both models were fit to each participant's classification ratings by identifying the parameters that minimized the squared difference between ratings and models' predictions. Parameter values averaged over participants are presented in Table 1. The table also presents for each model two measures of fit: the sum of squares error (*SSE*) averaged over participants and the

root mean square error (*RMSE*) averaged over participants¹. The conjunctive causes model showed a slightly better fit to the participants' ratings (lower *RMSE*). Consistently with regression analysis, in both models only parameters c_H and b_O showed probabilities considerably greater than 0.5, and all parameters representing causal relationships showed values near zero.

Average parameter values of Table 1 were used to generate the conjunctive causes model predicted values. These predicted values were then submitted to the same regression analysis as participant's ratings. The resulting predicted regression weights are presented in Figure 3 superimposed to participant's data. Note that the model nicely fits individual property's regression weights, but that even with parameters for causal relations as low as we found, the GM erroneously predicts higher coherence effects than those found in Experiment 1's results. This strengthens our conclusion that participants used individual properties to categorize, and did not use coherence.

3.3. Discussion

Results clearly show that participants were not categorizing by coherence. Regression coefficients were significantly different from zero for O, H and P, while there were no interactions between properties, as coherence would have predicted (the significant PO interaction involves a negative regression weight that would not be predicted by coherence). Model fitting supports the same conclusion. Model parameters suggest that participants weighed the presence of H and O at higher than chance levels, but did not take causal relationships into account to perform their ratings. These results conceptually replicate Hampton et al. (2009). However, as discussed earlier, coherence is not the only mechanism by which causal knowledge may influence categorization.

¹ The *RMSE* is a measure of fit that corrects for the different number of parameters in the two models. It's defined as $RMSE = \sqrt{SSE/(n - p)}$, where n = number of data points fit (32), and p = the model's number of parameters.

The next experiments were designed to test for the influence of prospective and retrospective inferences in artifact categorization.

Consistent with the process model proposed in Chaigneau, Barsalou and Sloman (2004), Experiment 1's results showed that O is an important property used to categorize artifacts. However, H was also considered by participants as a central property. This result is also consistent with results in Chaigneau, Barsalou and Sloman (2004) and many other studies in which H has shown its relevance for artifact conceptualization (e.g., Gelman & Bloom, 2000, Matan & Carey, 2001). We also explore this issue in the next experiments.

4. Experiment 2

At variance from Experiment 1, in this experiment participants did not find information about O. When confronted with important unknown information, people use available information to infer it. In prospective inference, a known cause is used to infer the state of its effect; in retrospective inference, a known effect is used to infer the state of its cause (see Fernbach, Darlow, & Sloman, 2011). And, these inferences can then be used to categorize (Rehder, 2010). Thus, depending on the causal relations participants perceive among H, P, G, A and O, prospective inferences predict that not providing information about O should introduce changes in the pattern of results from those found in Experiment 1.

As explained earlier, Figure 2 shows different causal models participants could perceive in the materials. Though participants did not show coherence effects, it's still possible that they would use causal knowledge to infer unobserved properties. Analyzing which properties were used in Experiment 1, allowed us to gauge which causal relations participants may use to make inferences in Experiment 2. Experiment 1's results suggest that participants did not conceive model (a) as the category's causal model because G and A did not influence their ratings. Though

this is a reasonable thing to expect (i.e., artifacts are not of a certain kind because an agent has a certain goal or uses them in a certain way, but because they afford the category's function), Experiment 2 allowed testing if this was a stable result for this set of experiments. In Experiment 1 only weights for P, H and O were significantly different from zero, which suggests that in Experiment 2 participants should consider, if any, only causal relations among those properties (i.e., a causal model similar to (b) or (c) in Figure 2). If participants in Experiment 2 perceive model (b), then we would expect them to use P to prospectively infer O's unknown state, making P gain relevance for categorization, while H would be screened-off (the causal Markov condition, see Pearl, 2000). In other words, model (b) implies that the distal cause H will become significantly less relevant when the proximal cause P is specified. However, model (c) is also possible. In the context of causal ascription, Lombrozo (2010) has proposed that people can treat human intentions (in this case, the designer's intention) as metaphorical mechanisms of causal transmission. In this sense, H could be viewed as a direct, albeit metaphorical, cause of an artifact's functional outcome, and could be used to infer it. Therefore, model (c) implies that Experiment 2 should produce a significant weight for H, which could be used to prospectively infer the state of O.

4.1. Method

4.1.1. Design and Participants

This experiment used a 3 (novel artifacts) x 16 (scenarios) mixed design. Twenty-four Adolfo Ibáñez undergraduates (native Spanish-speakers) participated in this study for course credit.

Participants signed informed consent and were randomly assigned to one of 3 artifacts and one of 4 random orders of scenarios.

4.1.2. Materials

Materials were equivalent to those of Experiment 1, except that information about the functional outcome was always absent. This meant that scenarios had 4 binary properties (H, G, A and P), and that participants provided $2^4 = 16$ ratings. The 16 scenarios for each category presented each participant with all combinations of intact and compromised H, G, A and P.

4.1.3. Procedure

Procedures were the same as in Experiment 1.

4.2. Results

4.2.1. Results for the regression analysis

To determine the centrality of properties, we analyzed participants' log transformed ratings by performing a multiple regression for each participant. Four predictor variables (H, P, G and A) were coded as -1 if the feature was compromised and +1 if it was intact. The regression weight associated with each predictor represents the influence that each element had on ratings. We also computed 11 interaction terms and introduced them in the individualized regression equations (6 two-way, 4 three-way and 1 four-way interactions). Averaged regression weights over participants for individual predictors and all interaction terms are presented in Figure 5.

Preliminary *t* tests showed that regression weights for H and P were significantly different from zero ($t(23) = 3.03, p < .01$; $t(23) = 4.83, p < .001$, respectively). Regression weights for interactions terms HA, PG, PA and HPGA were also significantly different from zero ($t(23) = 2.07, p < .05$; $t(23) = 2.10, p < .05$; $t(23) = 2.12, p < .05$; $t(23) = 2.30, p < .05$, respectively). Consequently, in this experiment ANOVAs were carried out for individual and interaction terms. These two ANOVAs were carried out separately for easier comparison between the regression weights of H, P, G and A.

Individual regression weights were entered in to a mixed 3 (artifact) x 4 (random scenario order) x 4 (property: H, P, G, A) ANOVA, with repeated measures on the last factor. There were no effects either of artifact or order and no interactions with the repeated measures factor and thus results were collapsed over these factors. A violation of the sphericity assumption was handled by correcting degrees of freedom with Huynh-Feldt's epsilon ($\epsilon = .617$). For clarity of presentation, degrees of freedom are presented here without adjustment. There was a main effect of property ($F(3, 69) = 9.36$, $MSe = .09$, $p < .001$, $R^2 = .29$, power = .96). Post hoc tests on the repeated measures factor (with Bonferroni adjustment), revealed that the regression weight associated with P was significantly greater than H, G and A (all $ps < .05$). There were no other significant differences.

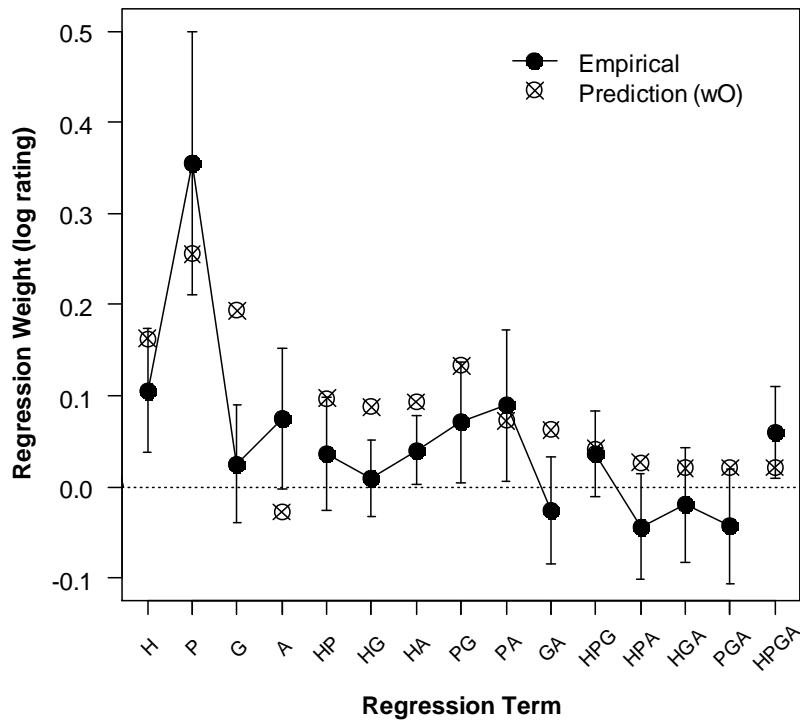


Figure 5. In Experiment 2, mean regression weights for history (H), agent goal (G), agent action (A), physical structure (P) and all possible interaction terms. Regression weights for H, P, HA, PG, PA and HPGA were significantly different from zero. Bars are 95% confidence intervals. Predictions derived from model fitting are shown superimposed on data.

Non-zero interaction regression weights were entered into a mixed 3 (artifact) x 4 (random scenario order) x 4 (term: HA, PG, PA, HPGA) ANOVA, with repeated measures on the last factor. There were no effects either of term, artifact or order and no interactions.

4.2.2. Model description

As in Experiment 1, we fit the GM to our data. For the purposes of this analysis, we used a model that is structurally equivalent to independent and conjunctive causes models in Figure 4 but that is without variable O and its corresponding causal links (wO model, see Figure 6). In the wO model we added a causal link from P to A because in this way model fitting would be able to retrieve all possible 2-way and higher interactions. Additionally, in previous experiments where information about O was missing (Chaigneau et al., 2004) participants did inform a causal link from P to A, which suggests us also including it here.

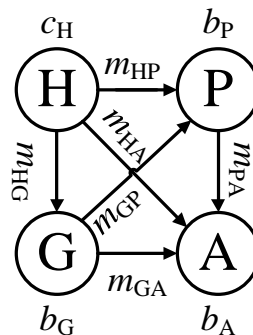


Figure 6. Causal model without variable O (wO model).

4.2.3. Model fitting

Following Experiment 1, we fit the wO model by predicting participants' log transformed ratings according to equation (3). Probability expressions for all exemplars were built applying equation (1). All free parameters were constrained as in Experiment 1.

Parameter values averaged over participants and measures of fit for the wO model are presented in Table 2. Parameters cH and bG showed values above chance, which is consistent with significant regression weights found for H and P. In addition, causal parameter mGP obtained an above zero value. (In contrast to measures of global fit, it is difficult to interpret individual causal parameters. Because of this, here and in the other experiments, we somewhat arbitrarily consider causal parameters above .2 to be indicative of a causal relation extracted by the model.) Figure 5 shows the wO model's predictions superimposed on Experiment 2 data. As can be seen, in general the model is not able to reproduce the correct pattern of weights for individual or interaction terms ($R^2 = .66$). We will discuss this in more detail later. However, consistently with the regression analysis, the model does predict coherence effects—but in a different pattern than revealed by data.

Table 2

Parameter estimates and measures of fit for the wO model (see Figure 6). Note that causal parameter mGP (in bold) suggests coherence effects. Also, base rate parameters for H and G (in bold) are above chance (0.5).

Parameters and measures	wO	
	M	SE
cH	0.591	0.026
bG	0.572	0.035
bP	0.553	0.037
bA	0.313	0.041
mHG	0.142	0.056
mHP	0.138	0.060
mHA	0.181	0.053
mGP	0.213	0.070
mGA	0.119	0.049
mPA	0.149	0.053
B	14.323	0.945
Average SSE	3.514	
Average RMSE	0.775	
R ²	0.661	

4.3. Discussion

As predicted by prospective inferences, P was the most relevant property for categorization. This result is consistent with the hypothesis that P is used to infer the state of O, and this explains its increased weight for categorization (i.e., greater than H, in contrast to the pattern found in Experiment 1). This may also explain why modeling fitted the data worse than in the previous experiment ($R^2 = .66$, in contrast to $R^2 = .81$ in Experiment 1). As shown in Figure 5, P produced a greater regression weight than predicted by the GM, presumably due to its inferential value, while H and G produced lower weights than estimated. In the GM, categorization depends on the likelihood of the pattern of present/absent properties, given the category's causal model. In consequence, it appears it cannot accommodate inferences to variables that are neither measured nor specified in the model (as happens with O in the current experiment). The next experiments will offer more evidence for this conclusion.

As in Experiment 1, H was also a relevant property for categorization (weight greater than zero). However, it showed a significantly lower weight than P (and not statistically different from G and A). This stands in contrast to the Experiment 1, where H was more central than P, and we think it's due to P's increased inferential value in Experiment 2. However, these comparisons across experiments are problematic and we need further experimentation to better understand this result.

In contrast with Experiment 1, participants' responses did show coherence effects. Interaction terms HA, PG, PA and HPGA were statically significant. This is in direct contradiction to Hampton et al.'s (2009) claim that artifacts lack coherent causal models. When information about the state of O was not provided, participants appear not only to have performed causally based prospective inferences, but also to use coherence to guide categorization. Though the following is

only hypothetical at this point, the reason for the interactions in this experiment may be that not having information about O triggered participants to use coherence to infer the state of O. We will test this explanation in the next experiments. Because we acknowledge that weights for interaction terms in this experiment are small (all $< .1$) and difficult to interpret without a control condition, in the next experiments we use a different procedure to test whether coherence effects increase when inferences are needed.

Relative to the effect of H in Experiment 2, it could be explained by models (b) or (c) in Figure 4. On the one hand, for model (b) the effect of H would imply a slight violation of the Markov condition (for other apparent violations, see Rehder & Burnett, 2005), but what is important here is that the general trend of the H mean coefficient being significantly lower than the P mean coefficient is consistent with participants using the $H \rightarrow P \rightarrow O$ causal chain model (i.e., when P's state is known, H's state has little relevance for knowing the state of O). On the other hand, the lower but still significant effect of H is easily accommodated by model (c), where H is viewed as a metaphorical, and thus perhaps weaker, cause of an artifact's functional outcome. Though Experiment 2's results are consistent with participants making prospective inferences from P to O and to a lesser extent from H to O, this issue needs further examination. This is the next experiments' main goal, which uses a single design to test for prospective and retrospective inferences between H, P and O. Also, these experiments will allow us to keep testing coherence effects and our tentative explanation for differences between experiments 1 and 2 in this regard.

5. Experiment 3

Because regression weights for G and A did not show significance in Experiments 1 and 2, in the current and following experiment they were treated as constant factors. In this way, scenarios in Experiment 3 contained information about H, P, G, A and O, but G and A were always intact. Aside

from reducing the number of interactions to analyze, this allowed participants to rate a smaller number of scenarios. Recall that in Experiment 1 participants rated a large number of scenarios. This may have led them to not focus on coherence and to use a simple categorization heuristic instead. In contrast, participants in Experiment 3 rated only 12 scenarios, reducing this concern.

In the current experiment, participants received scenarios where P and H could assume one of two values each (compromised or intact: -1, 1), while O could assume 3 values (compromised, absent or intact: -1, 0, 1). In consequence, 12 scenarios exhausted all possible combinations of H, P and O states. Our main goal for this experiment was to use a single design to know if people were making prospective inferences from H to O and from P to O. If this occurs, regression weights of H and P (as predictors of ratings), should increase when information about O is absent relative to when information about O is present. A second goal was to test again for interactions between properties that we found in Experiment 2 but no in Experiment 1. In particular, because in Experiment 3 H and P may assume 2 states (intact or compromised) but O may assume 3 states (intact, absent or compromised), this allowed analyzing if the HP interaction varied as a function of there being or not information about O. If our hypothesis that the need to make inferences about O triggers participants' use of causal coherence relations in categorization is correct (thus explaining the lack of coherence effects in Experiment 1 and their presence in Experiment 2), then the HP interaction term should increase when there is no information available about O.

5.1. Method

5.1.1. Design and Participants

This experiment used a 3 (novel artifacts) x 12 (scenarios) mixed design. Twenty-four Adolfo Ibáñez undergraduates (native Spanish-speakers) participated in this study for course credit.

Participants signed informed consent and were randomly assigned to one of 3 artifacts and one of 4 random orders of scenarios.

5.1.2. Materials

Materials were equivalent to those of Experiments 1 and 2, except that information about O was systematically manipulated assuming one of three values (-1, 0, 1), while P and H were binary properties (-1, 1). This manipulation resulted in $2 \times 2 \times 3 = 12$ scenarios.

5.1.3. Procedures

Procedures were the same as in the previous experiments.

5.2. Results

5.2.1. Results for the regression analysis

Recall that our data consisted on 3 predictors (H, P and O), where H and P were binary valued properties, and O could assume a third additional state (i.e., unknown). These three properties allowed computing 4 interaction terms, consisting on three 2-way interactions (HP, HO, PO), and one 3-way interaction (HPO). These individual and interaction terms were, as in previous experiments, entered into individualized regression equations and allowed us to measure the centrality of H, P and O, and to test for coherence effects.

To test for prospective inferences from P to O and from H to O, we computed an additional independent predictor labeled O^2 . Whenever information about O was provided (either 1 or -1), O^2 coded 1. Whenever the state of O was unknown, O^2 coded -2. The interaction term $P \times O^2$ represents the degree to which P's regression weight changes as a function of there being or not information about O. If positive, it reflects that the weight of P increases when there is information about O. If negative, it reflects that the weight of P increases when information about O is absent, as would be expected if participants performed prospective inferences. By analogous

reasoning, the interaction term $H*O^2$ represents the degree to which H's regression weight changes as a function of there being or not information about O, and allows testing prospective inferences from H to O.

Additionally, to test whether lacking information about O increased the use of coherence for categorization, we computed an additional independent HPO^2 interaction term. This interaction represents changes in the HP interaction weight as a function of there being or not information about O. A negative weight would show that the absence of O increases the HP interaction term's weight, as predicted by our triggering hypothesis.

Regarding prospective inferences from H to O and from P to O, two-tailed t tests showed that only the PO^2 interaction term was negative and significant ($t(23) = -3.59, p < .01$), consistent with the presence of prospective inferences from P to O. P's regression weight was .59 (95% CI [.27, .90]) when O was present compared to 1.19 (95% CI [.78, 1.60]) when O was absent (see Figure 8). In contrast, the HO^2 interaction weight was not significant ($t(23) = -1.69, p > .05$), thus not providing evidence of prospective inferences from H to O.

Regarding the HPO^2 interaction, it was not significant ($t(23) = .71, p > .05$), suggesting that participants did not engage in coherence based categorization, even when they needed to consider causal relations in order to make inferences about the state of O. Thus, this experiment did not support our triggering hypothesis.

Excluding the abovementioned interactions that involve superscripts, all other predictors were submitted to two-tailed t tests, with the null hypothesis that their mean regression weight was equal to zero. Only single properties H, P and O showed regression weights significantly different from zero ($t(23) = 3.37, p < .01$; $t(23) = 5.89, p < .001$; $t(23) = 6.18, p < .001$, respectively).

As in Experiment 1, when all information is provided, categorization ratings appear to be performed based on single properties, and not on coherence.

To compare the weights of single properties, their weights were entered in to a mixed 3 (artifact) x 4 (scenario order) x 3 (property: H, P, O) ANOVA, with repeated measures on the last factor. Results revealed a main effect of property ($F(2, 24) = 8.36$, $MSe = .06$, $p < .01$, $R^2 = .41$, power = .93), no property by order interaction ($F < 1$), no property by artifact interaction ($F(4, 24) = 1.47$, $MSe = .06$, $p > .05$, $R^2 = .20$, power = .39), and a significant three-way interaction ($F(12, 24) = 2.35$, $MSe = .06$, $p < .05$, $R^2 = .54$, power = .85). Post hoc tests were conducted on the repeated measures factor (with Bonferroni adjustment). This analysis showed that O's regression weight was higher and significantly different from the H's weight ($p < .01$), and that no other comparisons were significant (all $ps > .05$). The significant three-way interaction occurred because different orders of scenarios produced different weights for O for the tattoo-maker. To examine whether this interaction unduly affected our results, we repeated the analysis filtering-out the tattoo-maker. We found that the ANOVA produced the same pattern of statistical significance, and the post hoc difference between O's and H's weights remained (it was marginally significant with Bonferroni adjustment, $p = .065$; without adjustment, $p = .022$). Means for individual terms are shown in Figure 7.

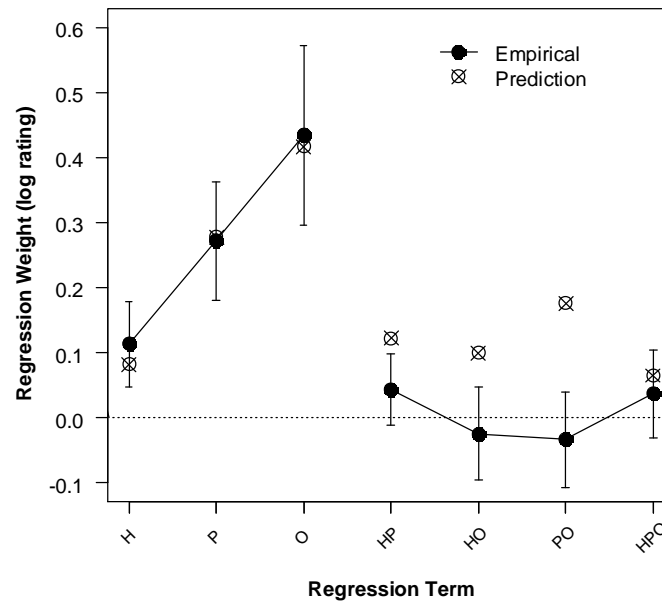


Figure 7. In Experiment 3, mean regression weights for history (H), physical structure (P), outcome (O) and all possible interaction terms. Only H, P and O are significantly different from zero. Error bars are 95% confidence intervals.

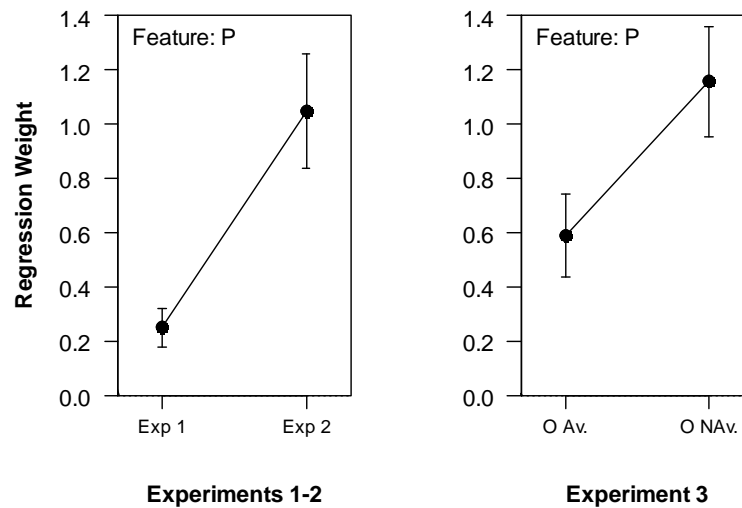


Figure 8. Difference between regression weights for P, when comparing Experiments 1 and 2 (left panel), and comparing conditions where information for O was available (O Av.) and not available (O NAv.) in Experiment 3. These graphs show regression weights without logarithmic transformation. Error bars are standard errors.

5.2.2. Model description

As in previous experiments, we fit the GM to our data. Because, as in Experiment 3, variables G and A were constant factors, we built a new model only including H, P and O (Figure 9). As with previous models, this model is able to represent, depending on which causal link is set to zero, both models (b) and (c) in Figure 2.

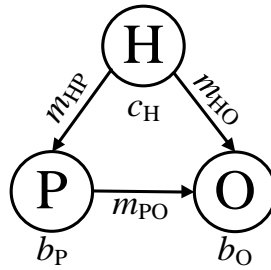


Figure 9. The H-P-O causal model.

5.2.3. Model fitting

Following previous experiments, we fit the H-P-O model by predicting participants' ratings according to equation (3). Probability expressions for all exemplars were built applying equation (1). All free parameters were constrained as in previous experiments.

Parameter values averaged over participants and measures of fit for the H-P-O model are presented in Table 3. Parameters b_P and b_O showed values above 0.5, which is consistent with significant regression weights obtained for all individual features. In addition, all parameters representing causal relationships showed values above zero. Figure 7 shows the H-P-O model's predictions superimposed on Experiment 3 data. As can be seen, while the model matches the pattern of weights for individual terms, it's not able to reproduce the correct pattern of weights for interaction terms ($R^2 = .36$).

Recall that our explanation for this lack of fit is that, as in Experiment 2, participants in the current experiment performed inferences and that the current versions of the GM is unable to accommodate inferences to variables that are neither measured nor specified in the model. To support this explanation, we repeated modeling for Experiment 3, but using only those exemplars where complete information was provided. This model fitting produced a substantial increase in fit (new $R^2 = .78$) and a pattern similar to that of Experiment 1 (a good fit for individual properties, and a consistent overestimation of coherence effects). We repeat this analysis strategy in Experiment 4 to support our conclusion the current GM implementation is unable to represent the inference to hidden variables.

Table 3

Parameter estimates and measures of fit of the H-P-O model. Base rate parameters for P and O (in bold) are above 0.5, and causal parameters (m_{ij}) are all above zero (also in bold).

Parameters and measures	M	SE
cH	0.544	0.023
bP	0.592	0.031
bO	0.698	0.053
mHP	0.261	0.078
mPO	0.288	0.088
mHO	0.340	0.091
B	5.562	0.376
Average SSE	4.709	
Average RMSE	0.940	
R2	0.357	

5.3. Discussion

Recall that our explanation for why P had a large weight in Experiment 2 appealed to the existence of prospective inferences. If participants conceived the scenarios as models (b) and (c) in Figure 2 illustrate, they may have used those relations to guide inferences about the state of the unknown O. The current experiment tested that explanation and found evidence consistent with

prospective inferences from P to O. When there was no information about O, participants appear to have used P to infer the state of O, increasing P's categorization value. However, we did not find evidence of prospective inferences from H to O. H did not significantly increase its weight when information about O was absent. This is interesting because it contradicts the idea in model (c) that H is conceived as a metaphorical cause of O. If it were, then participants could have used H to infer the state of O. However, because H was a property with relatively low weight, it may be that the lack of interactions involving H is due to lack of sensitivity of our experiment or due to floor effects.

Under complete information conditions, results replicate those of Experiment 1, with H, P and O correlating with categorization ratings. In the current experiment, O was the most central property, significantly more so than H, but not significantly different from P (again, a similar pattern as in Experiment 1). As in the first experiment, we failed to find evidence of participants using coherence to categorize—both under incomplete and complete information conditions. This questions our hypothesis that the need to perform inferences in Experiment 2 may account for the coherence effects we found. We return to this issue in Experiment 4.

6. Experiment 4

In Experiment 4, we tested prospective and retrospective inferences between H and P (as model (b) suggests) and retrospective inferences from O to P (as both, models (b) and (c) suggest). Relative to coherence effects—and considering contradictory results from experiments 2 and 3—in Experiment 4 we tested coherence again and also the hypothesis that the need to perform inferences may trigger the use of causal coherence in categorization.

In this experiment, G and A were again treated as constant factors. Scenarios in Experiment 4 contained information about H, P, G, A and O, but G and A were always intact. To

test for prospective and retrospective inferences between H and P (as model (b) in Figure 2 suggests), retrospective inferences from O to P (as both, models (b) and (c) suggest) and retrospective inferences from O to H (as model (c) suggests), participants received scenarios where P and H could assume three values each (compromised, absent or intact: -1, 0, 1), and O could assume only 2 values (compromised or intact: -1, 1). Consequently, participants needed to rate 18 scenarios.

6.1. Method

6.1.1. Design and Participants

This experiment used a 3 (novel artifacts) x 18 (scenarios) mixed design. Thirty-six Adolfo Ibáñez undergraduates (native Spanish-speakers) participated in this study for course credit. Participants signed informed consent and were randomly assigned to one of 3 artifacts and one of 4 random orders of scenarios.

6.1.2. Materials

Materials were equivalent to those of Experiment 3, except that information about H and P was systematically manipulated assuming one of three values (-1, 0, 1), while O was a binary property (-1, 1). This manipulation resulted in $3 \times 3 \times 2 = 18$ scenarios.

6.1.3. Procedures

Procedures were the same as in the previous experiments.

6.2. Results

6.2.1. Results for the regression analysis

Recall that our data consisted on 3 predictors (H, P and O), where H and P could assume any of three values, and O could assume only two states. These three properties allowed computing 4 interaction terms, consisting on three 2-way interactions (HP, HO, PO), and one 3-way interaction

(HPO). These individual and interaction terms were, as in previous experiments, entered into individualized regression equations and allowed us to measure the centrality of H, P and O, and to test for coherence effects.

To test for prospective and retrospective inferences between H and P and retrospective inferences from O to P, we computed additional independent predictors for H and P (labeled with a 2 superscript). Whenever information about H or P was provided (either 1 or -1), H^2 and P^2 coded 1. Whenever the state of H or P was unknown, H^2 and P^2 coded -2. The interaction term $P*H^2$ represents the degree in which P's regression weight varies as a function of there being or not information about H. This corresponds to retrospective inferences from P to H. The interaction term $H*P^2$ represents the degree in which H's regression weight varies as a function of there being or not information about P. This corresponds to prospective inferences from H to P. The interaction term $O*P^2$ represents the degree in which O's regression weight varies as a function of there being or not information about P. This corresponds to retrospective inferences from O to P. Finally, we computed an $O*H^2$ interaction term, representing retrospective inferences from O to H.

Note that, parallel to Experiment 3, H^2 and P^2 terms allowed computing interaction terms to test whether the absence of information about H or P interacted, respectively, with the magnitude of the PO and HO interactions. Negative POH^2 and HOP^2 interactions would show that, respectively, PO and HO interactions (indicative of coherence) increase when the state of H or P is unknown, as the triggering hypothesis predicts.

Ratings were, as in previous experiments, log transformed. Regarding prospective and retrospective inferences, two-tailed t tests showed that only the OP^2 interaction term was significant ($t(35) = -3.64, p < .001$), consistent with the presence of retrospective inferences from O

to P. O's regression weight was 1.37 (95% CI [1.11, 1.63]) when P was present compared to 1.76 (95% CI [1.50, 2.02]) when P was absent (see Figure 11). We found no evidence of prospective or retrospective inferences involving H: PH^2 ($t(35) = -.14, p > .05$); HP^2 ($t(35) = 1.27, p > .05$); and OH^2 ($t(35) = -0.96, p > .05$).

Consistently with our triggering hypothesis, which attempts to account for why we found coherence effects in Experiment 2 but not in Experiment 1, the POH^2 interaction was significant ($t(35) = -2.53, p < .05$) and the HOP^2 interaction was marginally significant ($t(35) = -1.90, p = .066$). These results are consistent with participants using coherence when information is lacking but not when it's complete.

Single properties and interaction terms not involving superscripts were submitted to a two-tailed t test, with the null hypothesis that the mean regression weight was equal to zero. Properties terms H, P and O showed regression weights significantly different from zero ($t(35) = 4.82, p < .001$; $t(35) = 5.48, p < .001$; $t(35) = 13.36, p < .001$, respectively). Also interaction term PO showed a regression weight significantly different from zero ($t(35) = -2.92, p < .01$). Note that this last interaction weight is negative and would not be predicted by coherence.

To compare the weights of single properties, their weights were entered in to a mixed 3 (artifact) x 4 (scenario order) x 3 (property: H, P, O) ANOVA, with repeated measures on the last factor. There were no effects either of artifact or order and no interactions with the repeated measures factor and thus results were collapsed over these factors. A violation of the sphericity assumption was handled by correcting degrees of freedom with Huynh-Feldt's epsilon ($\epsilon = .846$). For clarity of presentation, degrees of freedom are presented here without adjustment. There was a main effect of property ($F(2, 70) = 19.23, MSe = .08, p < .001, R^2 = .36, \text{power} > .99$). Post hoc tests were conducted on the repeated measures factor (with Bonferroni adjustment). This

analysis showed that O's regression weight was higher and significantly different from H's weight ($p < .001$), and also higher and significantly different from P's weight ($p < .001$). The difference between P and H was not significant.

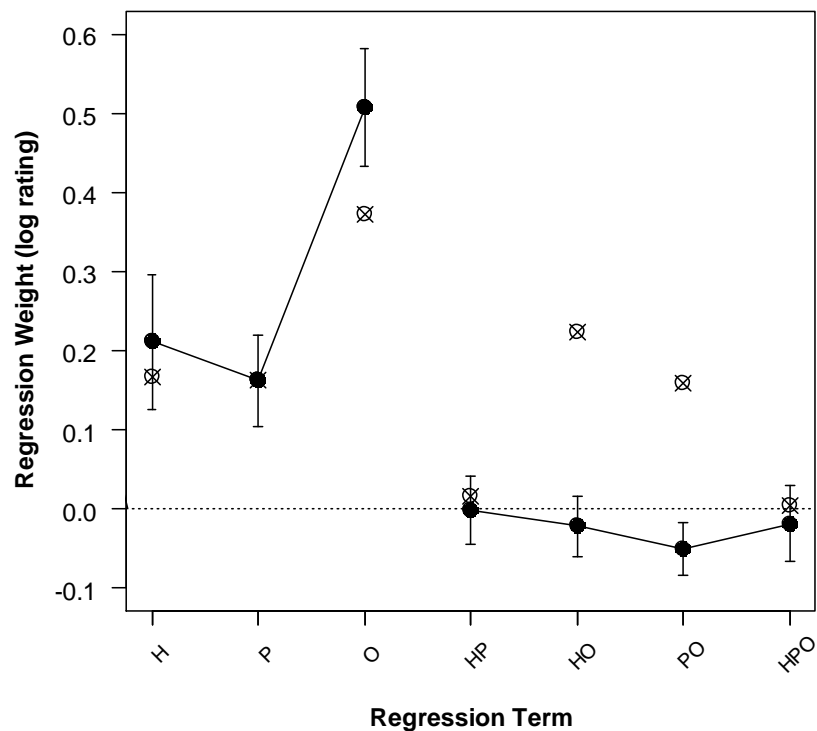


Figure 10. In Experiment 4, mean regression weights for history (H), physical structure (P), outcome (O) and all possible interaction terms. Weights for H, P, O and the PO interaction term are significantly different from zero. Error bars are 95% confidence intervals.

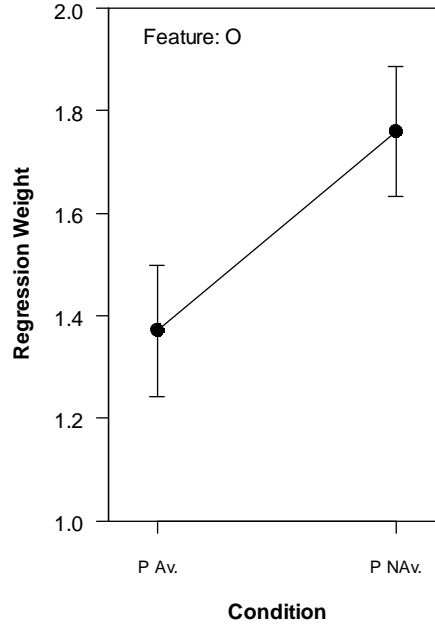


Figure 11. In Experiment 4, difference between regression weights for O, when comparing conditions where information for P was available (P Av.) and not available (P NAV.). The graph shows regression weights without logarithmic transformation. Error bars are standard errors.

6.2.2. Model description

As in previous experiments, we fit the GM of categorization to our data. Because in this experiment, H, P and O were the only variables, we used the H-P-O model (Figure 9).

6.2.3. Model fitting

Following previous experiments, we fit H-P-O model by predicting participants' ratings according to equation (3). Probability expressions for all exemplars were built applying equation (1). All free parameters were constrained as in previous experiments.

Parameter values averaged over participants and measures of fit for the three models are presented in Table 4. Parameters c_H and b_P (but not b_O) showed values above .5, which is partially consistent with significant regression weights found for all individual features. In fact, the model predicts a smaller regression weight for O than empirically obtained (see Figure 10). In

addition, causal parameters mPO and mHO showed values greater than zero (but not mHP). As seen in Figure 10, the model is not able to reproduce the correct pattern of weights for interaction terms ($R^2 = .38$). As in Experiments 1 and 3, the model predicts higher interaction weights than empirical ones.

Table 4. Parameter estimates and measures of fit of the H-P-O model. Note base rate parameters for H and P (in bold) are greater than .5 and that causal parameters (m_{ij}) mPO and mHO are greater than zero (in bold).

Parameters and measures	M	SE
cH	0.620	0.035
bP	0.629	0.030
bO	0.408	0.045
mHP	0.137	0.058
mPO	0.453	0.069
mHO	0.560	0.067
B	3.504	0.127
Average SSE	8.253	
Average RMSE	0.849	
R2	0.380	

To test again, as in Experiment 3, whether the GM is better able to fit data when information is complete, we performed our model fitting considering only exemplars with complete information. Consistently with our previous reasoning, when only complete scenarios were considered, fit substantially increased (new $R^2 = .62$), and the GM nicely predicted the individual property pattern, but consistently overestimated coherence effects. We will return to these issues in the General Discussion section.

6.3. Discussion

Relative to our question about causal inferences, the current experiment again showed that participants performed inferences to fill in absent information. In particular, they used information about O to retrospectively infer the state of P. Taking Experiments 3 and 4 jointly into

account, the $P \rightarrow O$ link appears stable across experiments, allowing prospective (Experiment 3) and retrospective inferences (the current experiment). In contrast, participants did not use H to make inferences and H did not receive inferences from other properties, which questions whether H was seen by participants as part of the overall causal model, and in particular, it questions model (c) in Figure 2 which posits a direct $H \rightarrow O$ link. However, given that H was a relatively weak feature in experiments 3 and 4, the last pattern of results might be reflecting a floor effect instead of an inferential independence of H respect P and O. In the general discussion we will come back to this issue by considering the behavior of H across experiments 1-4.

Experiment 4's results reconcile apparent contradictions from our previous experiments. When complete information is provided, results replicate those of experiments 1 and 3, with H, P and O being correlated with categorization ratings, and participants not giving evidence of using coherence to categorize. Under those conditions, in the current experiment O was again the most central property, with no significant differences between P and H. In contrast, when information about H or P was lacking, participants did increase the value of coherence for categorization (consistent with our triggering hypothesis). When H was lacking, the PO interaction increased; when P was missing, the HO interaction increased.

7. General Discussion

7.1. Artifact Causal Models

Returning now to our initial motivating question, our results show the need to qualify Hampton et al.'s (2009) statement. Replicating Hampton et al., when all information was provided (H, P, G, A and O) participants categorized artifacts based on individual properties and showed no evidence of using coherence for categorization. However, under incomplete information conditions, our results consistently show that participants were able to categorize artifacts based on causal

inferences, as evidenced by an increase of the regression weight of properties with inferential value (Experiments 2, 3 and 4). Inference draws on causal models resulting from the interaction of prior knowledge about artifacts and the structure of our materials, so it is questionable that—as Hampton et al. argue—artifacts lack strong coherent underlying causal models.

Regarding coherence, though our evidence is not conclusive (mainly due to null results for interactions in Experiment 3), we believe our experiments present preliminary support for the idea that the need to perform causal inferences highlighted causal links for participants, triggering their use of property coherence as a source of information for inferences and categorization. Based on our data, it's possible that coherent properties supported stronger inferences about unobserved properties than what would be expected from the linear combination of the individual pieces of information. More controlled experimentation may in the future resolve this issue.

7.2. Modeling and the Generative Model

Consistently across our experiments, the GM's ability (in its current implementation) to fit data decreased when participants used available information to infer absent properties. Best fits were always obtained when data for modeling included only exemplars with complete information. This is consistent with the view that, when our participants performed inferences about unobserved properties (as our regression analyses showed), the GM lacked means to take those hidden properties into consideration.

7.3. Relative Centrality of Individual Properties

There has been a fair amount of controversy about the relative importance of H, P and O for artifact categorization. In all our experiments, O was the most central property and other properties appear to have derived their weight for categorization partially from their value for inferring O. In particular, P became more central than O only when it allowed inferences about O

(as Experiment 3 showed). This provides an answer to controversies regarding whether P or O is more important. Such a controversy occurred between, e.g., Malt and Johnson (1992) and Barton and Komatsu (1989). In the former study, participants judged P as more important for categorization than O, while in the latter study participants viewed O as more important. In Malt and Johnson (1992, Experiment 2), objects were described that combined the standard physical properties of an artifact with functions varying in their degree of resemblance to the standard one (i.e., normal function, related function, bizarre function). For example, the standard physical properties for *boat* were *being wedge-shaped, having a sail, an anchor, and wooden sides*; whereas the normal function for *boat* was *transporting people over water*, the related function was *holding criminals off-shore*, and the bizarre function was *reintroducing marine animals to their habitat*. Participants in this experiment did not seem concerned about the non-standard functions, and granted category membership to at least half the items with unusual functions, which Malt and Johnson interpreted as meaning that P (the standard physical properties) was more central for categorization than O (the different functions). Our current data, which shows that O is more central than P, can explain these results by assuming that the Malt and Johnson participants granted category membership to artifacts with unusual functions because they inferred from their physical description that the artifacts were suited to achieve the standard function.

Aside from the P versus O controversy, there has also been a long H versus O controversy. In this controversy, some authors argue that H, and not O, is an artifact category's causal essence and therefore conceptually central (e.g., Ahn, Kalish, Gelman et al., 2001; Gelman & Bloom, 2000; Matan & Carey, 2001; see Strevens, 2000 for a critical view of essentialism). Our results fall squarely against this position. When participants were provided with complete information, H was never more central for categorization than O. In fact, O's weight was always nominally higher

than H's, significantly outweighing H in two out of three experiments (Experiments 3 and 4 versus Experiment 1). Though H was always correlated with categorization ratings, it did not behave as an essence. O was significantly more relevant.

Regarding this same issue, causal essentialism argues that an essence's centrality derives from its deep causal position in a category's causal model. Our data also comes out against this view as applied to artifacts. H—which would be artifacts' causal essence—was relevant for categorization (as discussed above) but our data provides scant evidence that it was part of participants' causal models. Judging from H's centrality in Experiment 1, where its weight was not different from O's weight, one might have predicted that it would allow many prospective and retrospective inferences. However, when participants lacked complete information, H did not show evidence of allowing or receiving inferences (Experiments 3 and 4) and only on a few occasions cohered with other properties (Experiments 2 and 4). P provides a point of comparison. P was involved in prospective and retrospective inferences (towards and from O, in Experiments 3 and 4) and cohered with other variables more often than H (with G and A in Experiment 2 and O in Experiment 4). Though P and O provided plenty of evidence of being linked in participants' models, H appeared at best weakly linked. In this context, it is difficult then to continue to argue that H's relative centrality derives from it being the category's deepest cause. H showed its centrality whenever participants were provided with complete information scenarios, but our experiments suggest this cannot be attributed to its causal role in participants' causal models. Our results suggest that some other non-causal factor may account for why in our experiments (in particular, Experiment 1)—and also outside the laboratory, H is viewed as conceptually central (see Bloom, 2007; Chaigneau, Castillo, & Martínez, 2008; Chaigneau & Puebla, 2013).

Acknowledgments

We are grateful to Cristián Coó, Vicente Soto and Mauricio Ríos for their help in data collection. We are also grateful to Bob Rehder for valuable comments and suggestions for the experiments and earlier versions of this manuscript. This work was supported by FONDECYT grant 1100426 to the second author. Parts of these analyses were presented at the 33rd Annual Meeting of the Cognitive Science Society.

References

- Ahn, W. (1998). Why are different features central for natural kinds and artifacts? *Cognition*, 69, 135-178.
- Ahn, W.K., Kalish, C., Gelman, S.A., Medin, D.L., Luhmann, C., Atran, S., Coley, J.D., & Shafto, P. (2001). Why essences are essential in the psychology of concepts. *Cognition*, 82(1), 59-69.
- Ahn, W., & Kim, N.S. (2001). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *The Psychology of Learning and Motivation*, Vol. 40 (pp. 23-65). San Diego, CA: Academic Press.
- Ahn, W., Kim, N.S., Lassaline, M.E., & Dennis, M.J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, 41, 361-416.
- Barton, M.E. & Komatsu, L.K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research*, 18(5), 433-447.
- Bloom, P. (2007). More than words: A reply to Malt and Sloman. *Cognition*, 105, 649-655.
- Chaigneau, S.E., Barsalou, L.W., & Sloman, S. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General*, 133, 601-625.
- Chaigneau, S.E., Castillo, R.D., y Martínez, L. (2008). Creators' intentions bias proper function independently from causal inferences. *Cognition*, 109, 123-132.
- Chaigneau, S.E., & Puebla, G. (2013). The Proper Function of Artifacts: Intentions, Conventions and Causal Inferences. *Review of Philosophy and Psychology*, 4(3), 391-406.
- Fernbach, P.M., Darlow, A., & Sloman, S.A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, 140(2), 168-185.

- Gelman, S.A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.
- Gelman, S.A., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition*, 76, 91-103.
- German, T., Truxaw, D., & Defeyter, M.A. (2007.) The Role of Information About "Convention," "Design," and "Goal" in Representing Artificial Kinds. *New directions for child and adolescent development*, 115, 69-81.
- Hampton, J. A., Storms, G., Simmons, C. L. & Heussen, D. (2009). Feature integration in natural language concepts. *Memory & Cognition*, 37(8), 1150-1163.
- Hernik, M. & Csibra, G. (2009). Functional understanding facilitates learning about tools in human children. *Current Opinion in Neurobiology*, 19, 34-38.
- Kelemen, D. & Carey, S. (2007). The essence of artifacts: Developing the design stance. In S. Laurence & E. Margolis (Eds.) *Creations of the Mind: Theories of artifacts and their representation*. Oxford: Oxford University Press.
- Kemler Nelson, D.G., Frankenfield, A., Morris, C., & Blair, E. (2000). Young children's use of functional information to categorize artifacts: three factors that matter. *Cognition*, 77, 133-168.
- Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Lombrozo, T. (2010). Causal-explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61, 303-332.

- Malt, B.C. & Johnson, E.C. (1992). Do artifact concepts have cores? *Journal of Memory and Language, 31*, 195-217.
- Malt, B.C. & Sloman, S.A. (2007). Artifact categorization: The good, the bad, and the ugly. In E. Margolis & S. Laurence (Eds.), *Creations of the Mind: Theories of Artifacts and Their Representation*. Oxford University Press.
- Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition, 78*, 1-26.
- Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.
- Minda, J.P., & Smith, J.D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 275-292.
- Nosofsky, R.M. (1984). Choice, similarity, and the context model of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104-144.
- Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.
- Oakes, L.M., & Madole, K.L. (2008). Function revisited: How infants construe functional features in their representation of objects. In R. Kail (Ed.), *Advances in child development and behavior, vol 36* (pp. 135-185). San Diego: Elsevier.
- Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.
- 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, 1141-1159.
- Rehder, B. (2010). Causal-based classification: A review. In B. Ross, (Ed.), *The Psychology of Learning and Motivation* (52), 39-116.
- Rehder, B. (2011). Reasoning with conjunctive causes. Paper presented at *The 52nd Annual Meeting of the Psychonomic Society*, Seattle, WA.
- Rehder, B. & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, 50, 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, 323-360.
- Rehder, B. & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 659-683.
- Rehder, B. & Kim, S. (2009). Classification as diagnostic reasoning. *Memory & Cognition*, 37, 715-729.
- Sloman, S.A., Love, B.C., & Ahn, W. (1998). Feature centrality and conceptual coherence, *Cognitive Science*, 22, 189-228.
- Strevens, M. (2000). The essentialist aspect of naïve theories. *Cognition* 74, 149-175.