Great apes and children infer causal relations from patterns of variation and covariation

Christoph J. Völter [a,b], Inés Sentís [c], Josep Call [a,b]

[a] Max Planck Institute for Evolutionary Anthropology, Department of Developmental and Comparative Psychology, Deutscher Platz 6, 04103 Leipzig, Germany

[b] University of St Andrews, School of Psychology and Neuroscience, South Street, St Andrews, Fife, KY16 9JP, Scotland, UK

[c] University of Barcelona, Faculty of Biology, Diagonal, 643, 08028 Barcelona, Spain

Correspondence concerning this article should be addressed to Christoph J. Völter, Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany, Email: christoph_voelter@eva.mpg.de, Phone: + 49 341 3550 450

Abstract

We investigated whether nonhuman great apes (N=23), 2.5-year-old (N=20), and 3-year-old children (N=40) infer causal relations from patterns of variation and covariation by adapting the *blicket detector* paradigm for apes. We presented chimpanzees (*Pan troglodytes*), bonobos (*Pan paniscus*), orangutans (*Pongo abelii*), gorillas (*Gorilla gorilla*), and children (*Homo sapiens*) with a novel reward dispenser, the blicket detector. The detector was activated by inserting specific (yet randomly determined) objects, the so-called *blickets*. Once activated a reward was released, accompanied by lights and a short tone. Participants were shown different patterns of variation and covariation between two different objects and the activation of the detector. When subsequently choosing between one of the two objects to activate the detector on their own all species, except gorillas (who failed the training), took these patterns of correlation into account. In particular, apes and 2.5-year-old children ignored objects whose effect on the detector completely depended on the presence of another object. Follow-up experiments explored whether the apes and children were also able to re-evaluate evidence retrospectively. Only children (3-year-olds in particular) were able to make such retrospective inferences about causal structures from observing the effects of the experimenter’s actions. Apes succeeded here only when they observed the effects of their own interventions. Together, this study provides evidence that apes, like young children, accurately infer causal structures from patterns of (co)variation and that they use this information to inform their own interventions.

Keywords: primate cognition, observational causal learning, problem-solving, blicket detector, comparative cognition

1 Introduction

A chimpanzee looking up at the canopy suddenly sees a group of colobus monkeys moving in the tree and feels simultaneously a gust of wind followed by a fruit falling to the ground (cf. Tomasello & Call, 1997). Based on this observation, the chimpanzee might learn associations between the presence of monkeys, the gust of wind, and the appearance of the fruit. Detecting such spatio-temporal associations in the environment is an essential step to make causal inferences about the world. However, mere associations even while taking into account important principles such as temporal priority or spatial contiguity are not always sufficient to infer causal structures (Hume, 1748 / 2000). For instance, based on the above observation alone, it remains ambiguous what caused the fruit to fall down (Seed & Call, 2009). One possibility is that the wind (W) and not the monkeys (M) caused the detachment of the fruit (F) (one-cause model: W → F). Alternatively, the gust of wind and the moving monkeys might be independent causes of a common effect (two-cause model: W → F ← M). Given the evidence, other models such as common cause and causal chain models are viable alternatives too.

Inferring causal structures in the environment based on the perceptual input is known as the *causal inverse problem* (Gopnik et al., 2004). Gopnik and colleagues proposed the differentiation of substantive and formal causal assumptions that might help an organism to solve this problem. On the one hand, substantive assumptions are specific causal principles such as the temporal order of cause and effect, spatial contiguity, and generally any prior knowledge about the world that constrains possible causal structures. On the other hand, formal assumptions provide a general, content-independent tool to infer causality-based patterns of correlation. These formal assumptions help us to distinguish between causal relations and mere correlations that are caused, for instance, by an unknown third factor (like an unobserved, common cause of two correlated variables).

Returning to the initial example, to resolve the aforementioned ambiguity between a one-cause model (W → F) and a two-cause model (W → F ← M), there are two options based on these formal causal assumptions: interventions and passive observations taking into account conditional probabilities of the events (Gopnik et al., 2004). First of all, intervening on each of the potential causes (e.g. chasing the monkeys away) while keeping the presence or absence of the other candidate cause constant would lead to different expectations depending on different causal structures. The second option is to observe situations in which only one of two co-occurring events is present. For instance, observing whether the monkeys' presence and the fruit's appearance are correlated depending on the presence of wind will reveal whether there is a relation between monkeys and the fruit's appearance independent of wind (as expected from a two-cause model but not from a one-cause model). The theoretical foundation for this is the causal Markov assumption (Hausman & Woodward, 1999) which states that given all direct causes of a variable are known and kept constant this variable will be independent of all other variables in the causal map except for its effects.

To shed light on infants' ability to learn about novel causal structures, in particular with regard to their ability to discount alternative candidate causes, Gopnik and colleagues (Gopnik & Sobel, 2000; Gopnik, Sobel, Schulz, & Glymour, 2001; Nazzi & Gopnik, 2000) developed a new experimental paradigm. Gopnik et al. (2001) presented 2.5- to 4-year-old children with a new device, the so-called *blicket detector*. This detector lit up and played a tune, if certain objects, the *blickets*, were placed on top of it. Other objects did not activate the blicket detector. Children were told that that blickets would always make the machine go. The task for the children was to identify objects that were "like blickets". The children received then different experimenter-given demonstrations. These demonstrations involved two novel objects but varied depending on the condition. In the one-cause condition, each object was placed on top of the detector by itself. One object (A) activated the detector; the other one (B) did not. Then both objects were placed on top of the detector simultaneously two times in a row and both times the detector was activated. In the two-cause condition, each object was placed on top of the detector by itself three times in a row. Whereas one object (A) activated the detector three times in a row, the other object (B) did not activate the detector the first time but did so the two following times. Thus, in both conditions one object (A) was associated with the activation of the detector in 100% of instances, while the other object (B) only in 67% of cases. However, in the one-cause condition, the effect of object B was conditional on object A. In contrast, in the two-cause condition the effect of object B on the detector was not conditional on A. Therefore, in the one-cause condition only object A could be like a blicket, whereas in the two-cause condition both objects might be regarded as blickets. Children's performance confirmed the hypothesized difference between the two conditions. In the two-cause condition, 3- and 4-year-old children were more likely to say that object B (the 67% object) was a blicket than in the one-cause condition. Moreover, in a forced-choice situation, 2.5-year-olds preferred object A over B in the one-cause condition but not the two-cause condition. Hence, Gopnik and colleagues (2001) concluded that young children infer novel causal relations by using conditional dependencies to discounting alternative candidate causes.

The extent to which the cognitive abilities of nonhuman great apes, our closest living relatives, might match those of humans is subject to ongoing debate. The relational reinterpretation hypothesis (Penn, Holyoak, & Povinelli, 2008) proposes that the cognitive differences between humans and nonhuman primates originate in the ability for abstract, relational reasoning. According to this view, nonhuman apes are incapable of re-interpreting perceptual input in terms of higher-order structural relations (e.g. reasoning about unobservable mechanisms and physical regularities). Contrary to this hypothesis, other scholars (Seed & Call, 2009) contended that nonhuman apes do have the capacity to encode and process information at an abstract, structural level, and not only at the perceptual level (allowing, for instance, for transferring knowledge between perceptually disparate but functionally equivalent tasks). In line with the latter view, there is some experimental evidence suggesting that great apes, at least in some situations, take unobservable object properties (such as weight and solidity) into account when solving problems (for recent reviews see, Seed & Call, 2009; Seed, Hanus, & Call, 2011).

Apart from this debate on nonhuman animals' ability to reason about unobservable causal mechanisms, a central question in this context is how nonhuman animals (as compared to humans) learn and represent novel causal structures. Penn and Povinelli (2007, p. 110) propose that "nonhuman animals' capacity for flexible goal-directed actions suggests that they explicitly represent the causal relation between their own action and its consequences". At least in the case of their own instrumental actions, nonhuman apes may be able to distinguish between covariation and causation. However, up to this point no study has explicitly addressed this issue, not to mention the question of whether apes are also able to distinguish between causation and covariation solely based on observational evidence (e.g. by observing others' interventions).

Under natural conditions, animals often face situations with multiple covarying events as alluded to in our opening example. In order to make efficient predictions about their environment animals would benefit from differentiating between causation and covariation. Causal discounting, or explaining away, is important to achieve this differentiation. Discounting means that the presence of one cause of an effect reduces the requirement of invoking other causes (Sloman, 2009). In certain situations, cue competition effects known from the associative learning literature can lead to similar outcomes. The nature of the cognitive processes underlying these cue competition or interaction effects is the subject of an ongoing debate (De Houwer, Vandorpe, & Beckers, 2005). Evidence for the involvement of inferential reasoning processes is provided by findings indicating that blocking effects are sensitive to ceiling effects and outcome additivity in rats and humans (Beckers, De Houwer, Pineno, & Miller, 2005; Beckers, Miller, De Houwer, & Urushihara, 2006). Additionally, the extent of training might be informative here. Cue competition effects in nonhuman animals are usually observed only after many exposures to the relevant contingencies (except for some specific contexts such as taste aversion). The demonstration of causal discounting after minimal exposure to the relevant contingencies (like in the blicket detector paradigm) would provide more evidence for the role of reasoning processes.

Compared to the literature on causal mechanisms, very few studies have examined the capacity of nonhuman primates to learn novel causal structures. One such study investigated whether nonhuman great apes (henceforth apes) were sensitive to the temporal order of cause and effect in the context of an object displacement task (Völter & Call, 2014). In this study, great apes needed to locate a yoghurt reward that was hidden under one out of two opaque cups and displaced out of their sight. Crucially, the yogurt baited cup left a yoghurt trail behind it. The apes spontaneously used the trail to locate the baited cup. Moreover, when presented with two perceptually identical trails leading to two different cups the apes ignored the trail that was already present before the cups were displaced and picked the cup at the endpoint of the causally relevant trail. This suggests that apes can integrate temporal information about cause and effect when making causal judgments.

The special role of causal interventions has been examined in nonprimate species. Blaisdell and colleagues (Blaisdell, Sawa, Leising, & Waldmann, 2006; Leising, Wong, Waldmann, & Blaisdell, 2008) provided evidence that rats (*Rattus norvegicus*) discriminate between the effects of their own interventions and observed effects. These studies suggest that when rats observe the effect of their own intervention they do not expect an alternative cause to be present. In contrast, when presented with the same effect that was not preceded by their intervention rats seemed to expect that an alternative cause was present. Thus, rats seem to use their own interventions in order to discount alternative causes. This does not mean necessarily that they also deliberately perform such interventions to begin with in order to elucidate causal structures (Penn & Povinelli, 2007).

Another study aiming at causal interventions compared New Caledonian crows (*Corvus moneduloides*), well-known for their habitual and flexible tool-use, with 2-year-old human children on a task in which subjects were required to insert an object into an apparatus (Taylor et al., 2014). The apparatus was made of transparent Plexiglas granting visual access to the bait and inner workings of the apparatus. If the object was inserted into the correct opening of the apparatus, a baited platform inside would revolve and dispense the reward. For crows observing the correlation between their own actions (i.e. “accidentally” inserting the pre-positioned object) and the appearance of the food reward was insufficient to produce the novel intervention. Only after a step-wise shaping procedure (involving on average more than 100 trials), they were able to pick up the object and insert it into the apparatus. Two-year-old children, in contrast, quickly produced the novel intervention after having observed the “accidental” intervention on average 3.5 times. Other studies, in contrast, found that egocentric experience with task-relevant actions (i.e. inserting stones in a tube) or mechanisms (a collapsing platform) was sufficient for crows to pick up objects and insert them into a different apparatus (Bird & Emery, 2009; von Bayern, Heathcote, Rutz, & Kacelnik, 2009) Therefore, it remains contentious whether the deficits in performance reported by Taylor and colleagues (2014) are evidence for a general shortcoming in crows’ ability to produce causal interventions or related to difficulties imposed by this particular task (Jacobs, von Bayern, Martin-Ordas, Rat-Fischer, & Osvath, 2015). One interesting suggestion put forward to facilitate the task by Taylor and colleagues (2014) was to remove the information about the causal mechanism of the apparatus. This would potentially help the birds to focus on the causal structure of the task (Jacobs et al., 2015). And it would effectively turn the apparatus into a blicket detector.

The blicket detector paradigm may be well-suited to investigate causal discounting in great apes (and other habitual tool-users) because it can be conceptualized as a tool selection paradigm. Great apes, in particular chimpanzees and orangutans, are exceptional among nonhuman primates with regard to tool-use and innovation rates (Lefebvre, Reader, & Sol, 2004). Tool selection studies have established that apes (as well as capuchin monkeys) take relevant properties of tools into account in order to select the most efficient tools. Their choice is flexible and dependent on the task they are facing (e.g., Manrique & Call, 2011; Manrique, Gross, & Call, 2010). The question arises how nonhuman apes initially acquire knowledge about relevant tools, tool properties, and actions and discount irrelevant ones.

Sensitivity to the difference between confounded (i.e. multiple variables are altered at once) and unconfounded interventions (i.e. only one variable is altered at a time) is relevant here. Chimpanzees, for example, learn over an extended period of time to use tools from observing their mothers as primary model (e.g., Biro et al., 2003). Discerning between confounded and unconfounded interventions seems relevant for them in order to learn task relevant contingencies. In contrast to other tool selection studies, in the blicket detector paradigm the task-relevant tool properties are completely opaque and arbitrary and cannot be deduced from physical knowledge or prior experience. The exclusive source of information regarding the functionality of a given objects are demonstrations by a human model.

In an attempt to adapt the blicket detector paradigm for nonhuman primates, Edwards and colleagues (2014) recently presented brown capuchin monkeys (*Cebus apella*) with a version of the blicket detector that delivered food when certain objects were placed on top of the device. They used one set of items per condition (one-cause and two-condition) and instead of demonstrations the monkeys could operate the detector on their own in the learning phase. The monkeys' performance did not differ between the two conditions as they preferred object A (i.e. the 100% object) over object B (i.e. the 67% object) in both conditions. However, when they were allowed to choose between the "B" objects (i.e. 67% objects) of the one-cause and two-cause conditions, the monkeys preferred the object belonging to the two-cause condition. Several important changes to the original version of the blicket detector paradigm make these results hardly comparable to children's performance, including the egocentric experience the monkeys got with the objects before the test phase (as compared to non-egocentric, experimenter-given demonstration in Gopnik and colleagues' version of the blicket detector, see Gopnik et al., 2001), the amount of experience they got with each step of the demonstration (a whole session for the monkeys compared to one or two demonstrations for the children), and the number of different sets of objects (one set for the monkeys vs. two sets of objects per condition for the children).

In the current study, we examined whether apes, like children, were able to identify the most likely cause by discounting alternatives when presented with patterns of variation and covariation. One of the main strengths of this study is its comparative dimension, more so for a task that has been argued to measure a fundamental component for human thought. It is important to ascertain whether this is unique to humans or a shared trait with our closest living relatives. Therefore, we adapted the original blicket detector paradigm for apes while maintaining its basic features including the type and amount of experience with the objects during the demonstration-phase and the number of trials per condition. A procedural adaptation of the task for apes was to use a forced-choice measure; like in Gopnik et al.'s study with 2.5-year-olds (2001, experiment 2). We sought to replicate the findings by Gopnik et al. with 2.5-year-olds and explored how 3-year-olds would perform with this forced-choice measure. Given similar performance of apes and 2.5-year-olds in different tasks aiming at causal and physical cognition (Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007) we expected similar performance between apes and 2.5-year-olds. The major advantage of this set-up was that the causal relations involved in the task were novel and arbitrary and we were able to control the kind of evidence that subjects received. In addition, to be able to better compare, we examined to what extent the performance of preschool children depends on verbal scaffolding.

2 Experiment 1

2.1 Material and Methods

2.1.1 Subjects. Twenty-one chimpanzees (*Pan troglodytes*), eight bonobos (*Pan paniscus*), five orangutans (*Pongo abelii*), four gorillas (*Gorilla gorilla*), 25 2.5-year-old, and 46 3-year-old human children participated in this experiment (see supplementary material for detailed information on the samples).

2.1.2 Materials. The blicket detector consisted of a grey box (30 x 42 cm) with an L-shaped, transparent Plexiglas tube (diameter 7cm, 12 x 14 cm) attached on top that could be turned either away from the subjects or toward the subjects. Via this tube, objects could be inserted into the apparatus. Inserted objects fell down into a chamber inside the apparatus (see Figure 1). The subjects could see what was inside this chamber as the front side of the chamber was made out of transparent Plexiglas (16 x 16 cm). Underneath the chamber, there was a hole (diameter 6 cm) in the box where the food was delivered upon activation of the detector. There were three, white LEDs (diameter: 2.3 cm, 12 V), one integrated in the left side wall of the chamber, the other two were integrated in the front side of the apparatus. The only modification that we introduced for the children was a yellow cardboard cover that we glued to the front side of the apparatus and that covered the grey parts of the apparatus to make it more appealing for the children.

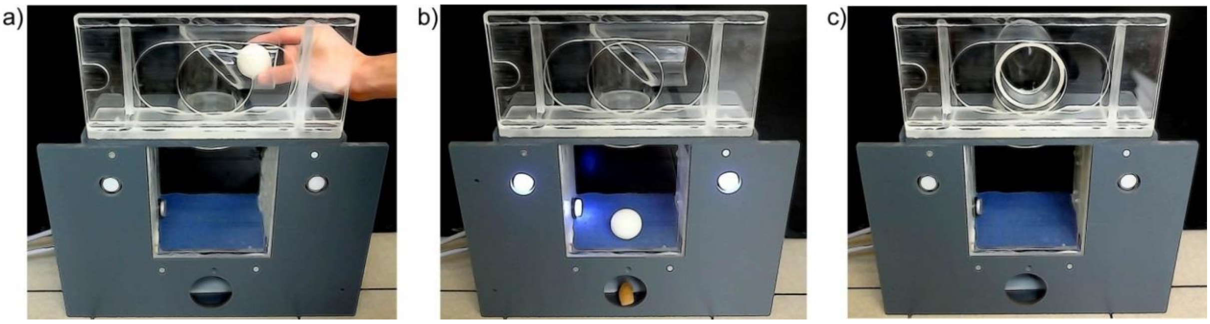


Figure 1. Illustration of the initial training phase, a) the experimenter inserts the white ball in the blicket detector via the tube on top of the apparatus, b) as soon as the ball touches the bottom of the chamber inside the apparatus three LEDs light up, the buzzer rings out, and the food appears simultaneously in the hole underneath the chamber, c) the experimenter turns the L-shaped tube on top of the apparatus towards the subject.

Out of the subjects' view, there was a hatch inside the apparatus on top of which the reward was placed. The hatch was controlled by an electromagnet (diameter 20mm, 15mm, 60N, 12V). Moreover, there was a buzzer (75 dB, 400 Hz, 12V) inside the apparatus that served as acoustic signal. The experimenter (E) could control the detector via a foot pedal that was hidden behind a screen. When E pressed the pedal the detector switched on, i.e. the three LEDs illuminated, the buzzer rang out, and the hatch with the reward was released. The released food (a dry food pellet for the apes and colored, wooden beads for children) rolled down a ramp and stopped in front of the food hole where the subject could see and access it. We used 21 objects of different shapes, colors, and materials. No two objects were identical. The objects were divided into pairs based on their overall size. The same objects were used for apes and children.

2.1.3 Procedure and Design. The general procedure was identical in every condition. For the apes, we fixed the blicket detector to the mesh of the test enclosure; for children we placed the blicket detector on a small table. Subjects had access to the food hole and the tube opening when the tube was turned toward the subject. Next to the blicket detector, we presented the two objects on a table that we positioned perpendicularly to the detector. A Plexiglas panel with two small, circular holes (6cm) on opposite sides prevented subjects from grabbing the objects. In the beginning of each trial, E placed two objects (which were completely novel to the subjects) on top of the table facing the subject. Each object was shown individually to the subject and then placed in front of either the left or the right hole in the panel. Next the subjects received a demonstration that varied across conditions (see below). After this demonstration, E turned the tube on top of the detector toward the subject and pushed the two objects simultaneously towards the two holes in the panel. Subjects could choose one object by sticking their fingers through one of the two holes. Subjects received the chosen object and could then insert it inside the apparatus. If they did not do so within ten minutes, we stopped the session and repeated this trial in the next session (two subjects during the test phase, subjects did not alter their choice when the trial was repeated). If subjects gave the object back through the hole in the panel through which they had received it, they got to choose again (N=1). A trial was only scored as successful if subjects inserted the object in the apparatus which was the case in every trial except the ones mentioned above.

2.1.3.1 Training phase. In the beginning of the first session, we familiarized subjects with the task. Therefore, we presented subjects with a white, wooden ball. E inserted the ball in the blicket detector that was switched on (see Figure 1). After the subject had retrieved the delivered reward, E placed the ball on the table, turned the tube on top of the detector toward the subject, re-baited the apparatus. After the subject had chosen the ball (the only available object), E gave the ball to the subject who could then insert it into the apparatus. After having passed this initial familiarization procedure (one trial), apes and children received in total four and two trials of discriminatory training, respectively (two trials per session). In every trial two novel objects were used. The location of the object on the platform was counterbalanced within each subject. The demonstration was as follows: each object was inserted by itself, one after the other in counterbalanced order. One object (randomly determined and counterbalanced across subjects) activated the detector (object A); the other one (object B) did not (A+, B- or B-, A+). If subjects chose the correct item, E went on with the next set of objects. If the subject chose the incorrect object, E repeated the demonstration (with the same set of objects) and allowed the subject to choose again. If they again chose the incorrect object, E repeated this procedure one more time. The criterion for passing this training phase was choosing the correct object in three out of four trials (apes) / two of two trials (children) within a maximum of three attempts.

2.1.3.2 Test phase. In the test phase, subjects received three different conditions of two trials each: baseline, one-cause, and two-cause condition (see Table 1). The baseline condition was identical to the training phase (A+, B- or B-, A+). However, subjects did not get the chance to correct themselves when they chose the incorrect object.

Table 1 *Overview on the different conditions administered in experiment 1 - 3*

Exp.	Condition	Demonstration	Expected Preference for A
1	One-cause	A+, B-, AB+, AB+	One-cause > Two-cause
	Two-cause	A+, A+, A+, B-, B+, B+	
2/3	Retrospective discounting	CD+, CD+, D-	Retrospective discounting > Backward blocking
	Backward blocking	CD+, CD+, C+	

Note. The conditions are shown together with type of demonstration that subjects got before they could choose between objects A and B, and the relative preference for object A that we expected between the conditions. (+) represents activation, (-) no activation of the detector.

In the one-cause condition (see supplementary video 1), E presented subjects again with two new objects and inserted them by themselves one after the other (in counterbalanced order). One object (A) activated the detector; the other one did not (B). Then E inserted both objects together (AB) two times in a row and both times the detector was activated. Thus, the demonstration was as follows: A+, B-, AB+, AB+ (or B-, A+, AB+, AB+). Thereafter, the subjects were allowed to choose between object A (100% object) and B (67% object).

In the two-cause condition, E presented subjects with two new objects. E inserted each object (A / B) by itself three times in a row. One object activated the detector all three times (object A), the other object did not activate the detector the first time but it did activate the detector the following two times (object B). The order of presentation was counterbalanced. Thus, the demonstration was as follows: A+, A+, A+, B-, B+, B+ (or B-, B+, B+, A+, A+, A+). Again, the subjects were then allowed to choose between object A (100% object) and B (67% object).

Apes received two test sessions with one trial per condition. Children received one test session with two trials per condition. In each trial, apes and children faced a novel set of objects. The order of conditions was pseudo-randomized with the restriction that the baseline condition was never the first condition which we administered.

Additionally, we examined the impact of verbal scaffolding on children’s performance. Therefore, we manipulated the type of verbal instruction given to 3-year-old children. One group received verbal scaffolding about the causal nature of the task following the procedure by Gopnik et al. (2001). E told the children that the machine was a "blicket machine" and that "blickets make the machine go." After the demonstration the experimenter presented the children with the choice between the two objects by saying: “Now it’s your turn. Which of the two objects is like the blicket?”. The other group of 3-year-olds as well as all 2.5-year-olds received no such verbal scaffolding: Children were introduced to the blicket detector by being told that the machine was a novel game. After the demonstration the E presented the children with the choice between the two objects by saying: “Now it’s your turn. Which of the two objects do you want to try?”.

Differences in the procedure between apes and children (adjustments to the different testing environments) affected the number of training trials and the number of trials per session. In order to complete experiment 1 within one session, we only administered two training trials with children. Consequently, we also adjusted our criterion to the number of training trials.

2.1.4 Scoring and analysis. We videotaped all sessions. We scored which object subjects inserted in the apparatus. A second coder scored 20% of all trials of the apes to assess interobserver reliability which was excellent ($K=1$, $N=30$, $p<0.001$).

We used a Generalized Linear Mixed Model (GLMM; Baayen, 2008) with binomial error structure and logit link function to analyze the effects of condition and species on the percentage of trials object A was inserted in the blicket detector. We included condition, group, and the side of the 100% object as fixed effects, the interaction between condition and group, and subject as random effects. To keep type I error rate at the nominal level of 5% (Barr, Levy, Scheepers, & Tily, 2013; Schielzeth & Forstmeier, 2009) we included all possible random slopes components (condition and side of the 100% object within subject) and also the respective correlations between random slopes and intercepts. As an overall test of the effect of the predictor variables we compared the full model with a null model lacking the fixed effects condition and species but comprising the same control predictors and random effects structure as the full model (Forstmeier & Schielzeth, 2011) using a likelihood ratio test (Dobson, 2002). P values for the individual effects were based on likelihood ratio tests comparing the full with respective reduced models (Barr et al., 2013; R function drop1).

We assessed model stability by comparing the estimates derived by a model based on all data with those obtained from models with the levels of the random effects excluded one at a time. This revealed the model to be stable with regard to the effects of condition, species, side of the 100% object, and session. Overdispersion appeared to be no issue (dispersion parameter: 0.83).

We used one-sample t-tests to test against the hypothetical chance level (p=0.5), binomial tests to test subjects' first trial performance in session 1 against the chance level, and Fisher's exact test to compare the first trial performance across conditions. All p-values reported here are exact and two-tailed (except for Fisher's exact test).

2.2 Results

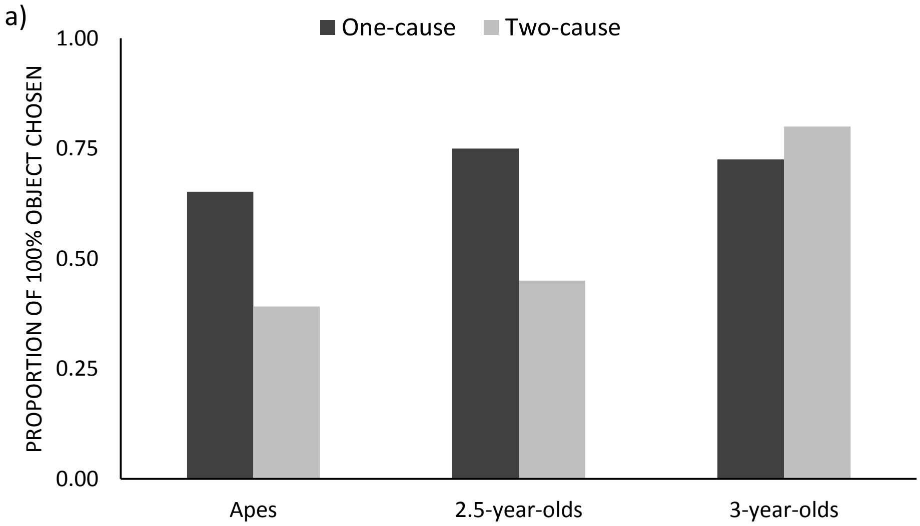
2.2.1 Training. Three out of eight bonobos (37.5%), seven out of 21 chimpanzees (33.3%), one out of five orangutans (20.0%), all four gorillas (100%), five of 25 2.5-year-olds (20%), and six of 46 3-year-olds (13%, four in the scaffolding condition and two in the demonstration-only condition) did not meet the training criterion as they either completely failed to insert the objects in the apparatus or failed to correct wrong choices. These subjects were excluded from all subsequent analyses. The remaining apes (N=23) scored significantly better than the expected chance value of 50% with regard to their first choice of the 100% object in the second training session (M±SE = 65.2 ± 6.6%; t(22)=2.30, p=0.031), but not in the first one (M±SE = 52.2 ± 6.7%; t(22)=0.33, p>0.250). Likewise, the remaining 2.5-year-olds (M±SE = 67.5 ± 7.5%; N=20; t(19)=2.33, p=0.031) and 3-year-olds (M±SE = 87.5 ± 3.5%; N=40; *t*(39)=10.82, *P*<0.001) scored significantly better than expected by chance in their training trials.

2.2.2 Test. We first analyzed whether there was a significant difference between the different species of nonhuman apes in our sample but neither the interaction between species and condition, or the main effect of species turned out to be significant (see supplementary material). Therefore, we collapsed the data of the different ape species in subsequent analyses. Moreover, we analyzed whether type of instruction influenced performance of the 3-year-olds but we did not find a significant interaction between instruction and condition or a significant main effect of instruction (see supplementary material). Therefore, we collapsed the data of 3-year-olds who received different types of verbal instruction in subsequent analyses.

Overall, the GLMM with group (apes, 2.5-year-olds, and 3-year-olds), condition (one-cause, two-cause, and baseline), the interaction of group and condition, and the location of the 100% object (left or right) was significant compared to the null model (likelihood ratio test: χ^2=28.10, df=8, p<0.001, see Figure 2). More specifically, we found a significant interaction between group and condition (χ^2=17.25, df=4, p=0.002). Pairwise comparisons revealed that 3-year-olds performed significantly better in the two-cause condition compared to 2.5-year-olds (z=2.65, p=0.008) and apes (z=4.50, p<0.001). We found no significant difference between 2.5-year-olds and the apes in the two-cause condition (z=1.67, p=0.096). In contrast, we found no differences between the groups in the baseline condition (3-year-olds vs. 2.5-year-olds: z=1.08, p>0.250; 3-year-olds vs. apes: z=0.18, p>0.250; 2.5-year-olds vs. apes: z=1.12, p>0.250) or the one-cause condition (3-year-olds vs. 2.5-year-olds: z=0.27, p>0.250; 3-year-olds vs. apes: z=0.22, p>0.250; 2.5-year-olds vs. apes: z=0.05, p>0.250). The location of the 100% object did not have a significant effect on performance (χ^2=2.59, df=1, p=0.108).

For apes the condition had a significant effect on performance (χ^2=8.59, df=2, p=0.014) with apes choosing object A more frequently in the one-cause (z=2.61, p=0.009) and baseline condition (z=2.39, p=0.017) compared to the two-cause condition. In contrast, apes showed no difference between the one-cause and baseline condition (z=0.27, p>0.250). For 2.5-year-olds (χ^2=5.49, df=2, *p*=0.064) or 3-year-olds (χ^2=4.93, df=2, *p*=0.085) there was no significant effect of condition on their performance.

In the one-cause condition apes (t(22)=2.31, p=0.030) and 3-year-olds (*t*(39)=2.76, p=0.009) but not 2.5-year-olds (t(19)=1.561, p=0.135) chose the 100% objects significantly more often than expected by chance. In contrast, in the two-cause condition apes tended to prefer the 67% object (t(22)=2.02, p=0.056). Whereas 2.5-year-olds (t(19)=0.81, p>0.250) did not show a preference for the 100% in the two-cause condition either, 3-year-olds (*t*(39)=7.26, p<0.001) chose the 100% object significantly more often than expected by chance also in the two-cause condition.



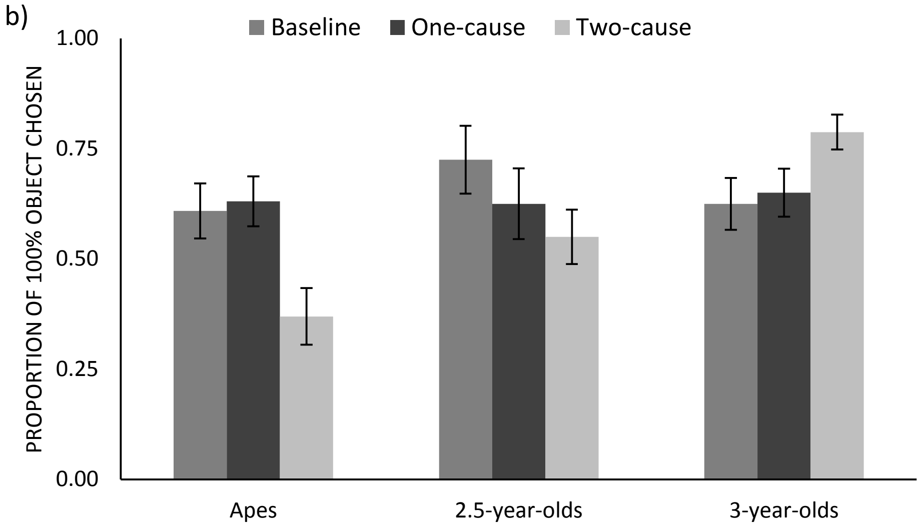


Figure 2. Exp. 1: a) Proportion (and number) of subjects who chose object A in their first trial across conditions; b) Proportion of trials in which object A (the 100% object) was selected across the different test conditions (mean ± SE). For each condition, we administered two trials per subject. In each trial, we presented subjects with a novel set of objects. Demonstration: baseline A+, B-; one-cause A+, B-, AB+, AB+; two-cause A+, A+, A+, B-, B+, B+.

To exclude carry-over effects across trials or sessions, we also analyzed whether the participants performed differently between the first and second trial of each condition. We found a significant interaction between trial number and condition for 2.5-year-olds (χ^2=6.54, df=2, *p*=0.038) but not for apes (χ^2=4.01, df=2, *p*=0.134) or 3-year-olds (χ^2=1.69, df=2, *p*=0.430; see Fig. 2a). Condition had a significant effect on 2.5-year-olds' performance in the first (χ^2=11.22, df=2, *p*=0.004) but not the second trial (χ^2=3.15, df=2, *p*=0.207). In their first trial, 2.5-year-olds performed significantly better in the one-cause and baseline condition (z=3.23, *p*=0.001) compared to the two-cause condition (z=2.74, *p*=0.006). We found no difference between the one-cause and baseline condition (z=0.11, *p*=0.914). Moreover, we analyzed separately the first trial of the first session. Apes (10 of 11 chose object A, 90.9%, binomial test: p=0.012), 2.5.-year-olds (9 of 11 chose object A, 81.8%, binomial test:

p=0.065), and 3-year-olds (18 of 21 chose object A, 85.7%, binomial test: *P*=0.002) showed a preference for the 100% object in the first trial of the one-cause condition. In the two-cause condition apes (6 of 12 chose object A, 50.0%) and 2.5-year-olds (1 of 9 chose object A, 11.1%, binomial test: p=0.039) did not exhibit a preference for the 100% object. In contrast, 3-year-olds also showed a significant preference for the 100% object in the two-cause condition (18 of 19 children chose object A, 94.7%, binomial test: p<0.001). In line with our hypothesis, apes (Fisher's exact test: p=0.045, one-tailed) and 2.5-year-olds (p=0.003, one-tailed) decided significantly more often for the 100% object in the one-cause compared to the two-cause condition. First trial performance of three-year-olds, in contrast, was not affected by condition (p>0.250).

2.3 Discussion

First, our findings demonstrate extremely fast discriminatory learning in apes and young children. After only a single demonstration with each object, apes and children were able to discriminate between two novel objects (apes in their second training session and children in their first session). Second, we found that the apes and 2.5-year olds discriminated between the one-cause and two-cause conditions. In line with our hypothesis, apes selected the 100% object (A) more often in the one-cause condition than in the two-cause condition and more often than expected by chance. Notably, this was already the case in the very first trial of session 1, thus, ruling out potential learning effects across trials or sessions. A similar pattern emerged for 2.5-year-old children in their very first trial: 2.5-year-olds, like the apes, preferred the 100% object over the 67% object in their first trial in the one-cause condition but not in the two-cause condition. Overall, the results for 2.5-year-olds did not differ significantly from the performance of the apes. Three-year-olds, in contrast, chose the objects that were consistently associated with the effect in all conditions, irrespective of the instruction they received, and they did so from the very first trial onwards. In contrast to apes and 2.5-year-olds, 3-year-olds preferred the 100% object also in the two-cause condition and they performed significantly better than apes and 2.5-year-olds in this condition.

Thus, apes and 2.5-year-olds preferred objects associated with the effect but only if they were not screened off by another object. These findings suggest that apes and 2.5-year-olds use conditional dependencies when making causal inferences in a completely novel and arbitrary context. Importantly, unlike a previous study with capuchin monkeys (Edwards, Rottman, & Santos, 2011; Edwards et al., 2014), the experience that the subjects got before they chose between the two objects was not egocentric, i.e. they were not allowed to try the objects out by themselves but had to observe an experimenter-given demonstration. Based on this observational information, apes and children produced their own interventions on the task.

We did not find any significant species differences among nonhuman apes. However, given the small sample size per species this negative result is unsurprising. We note that gorillas failed the training altogether. Three out of four gorillas did not even try to insert the initial object into the apparatus within four sessions after having observed a demonstration. A greater level of neophobia relative to other ape species might have contributed to their failure, especially given that we did not habituate subjects to the objects before the test.

Could procedural differences between children and apes, in particular, with regard to the number of training trials and trials per session, account for the present findings? We think this is unlikely for two reasons. First, apes and 2.5-year-olds performed at comparable levels in their final two training trials before entering the test phase. Second, the first trial analysis showed a similar pattern for apes and 2.5-year-olds (see Figure 2a). A similar training outcome might be the best way to ensure a valid comparison between children and apes given their different experience with human artifacts.

In contrast to the original study by Gopnik and colleagues (2001, Experiment 1), we presented 3-year-olds here with a forced-choice test. Three-year-olds in Gopnik et al. study performed at ceiling when asked whether object A was like a blicket. Therefore, it is not surprising that three-year-olds in the current study preferred the most effective object in both the one-cause and two-cause condition. This finding is also in line with results by Kushnir and Gopnik (2005) showing that 4-year-olds prefer objects (in a forced-choice setting) that are more effective in activating the blicket detector by itself. More specifically, they prefer objects that are associated with the effect in 3/3 vs 1/3 instances or 2/3 vs 1/3 instances. Our study serves to extend these findings to 3-year-olds who preferred more effective objects (3/3) over less effective objects (2/3).

In the current study, the 2.5-year-olds performed better in the one-cause condition compared to the two-cause condition but only in their first trial in each condition. These differences might be due to changes in the experimental procedure compared to the original study. First, in the original study blickets were not only associated with the onset of the detector but removal of the blickets was also associated with the termination of the effect. In the current study, the insertion of blickets activated the detector; however, the effect ended by itself after a fixed period of time. Therefore, only the onset of the effect but not its termination was contingent upon the blicket which might have resulted in a less salient relation between blicket and detector. Moreover, in the current setup the activation of the detector included dispensing a reward, which might have distracted them from the relevant contingencies or changed their motivation to maximize the number of rewards by “playing it safe” (i.e. choosing the 100% throughout the different conditions).

Despite these differences in design, we found evidence that in a forced-choice procedure apes, like 2.5-year-olds, used patterns of variation and covariation to infer accurately novel causal structures. Three-year-olds, in contrast, performed rationally by choosing the most effective objects throughout the different conditions. Next, we investigated whether apes and children would also be sensitive to a more complex type of conditional independence involving retrospective re-evaluation of the evidence.

3 Experiment 2

3.1 Material and Methods

3.1.1 Subjects. Sixteen chimpanzees, six bonobos, four orangutans, 15 2.5-year olds, and 36 3-year-olds participated in this experiment. All subjects had participated in Experiment 1 before. The interval between the experiment 1 and 2 was on average 43 days (range: 30-50 days) for the apes, 2 days (1-4 days) for 2.5-year-olds, and 3 days (1-13 days) for 3-year-olds.

3.1.2 Materials. The apparatus was the same as in Experiment 1. Twelve novel objects were used as stimuli.

3.1.3 Procedure and Design. The basic procedure including the randomization was identical to Experiment 1. Subjects that passed the training phase in Experiment 1 did not get any more training trials in the current experiment. Apes that failed the training in Experiment 1 got another two sessions of training as in Experiment 1 but with novel objects. The same criterion for entering the test phase was used as in Experiment 1

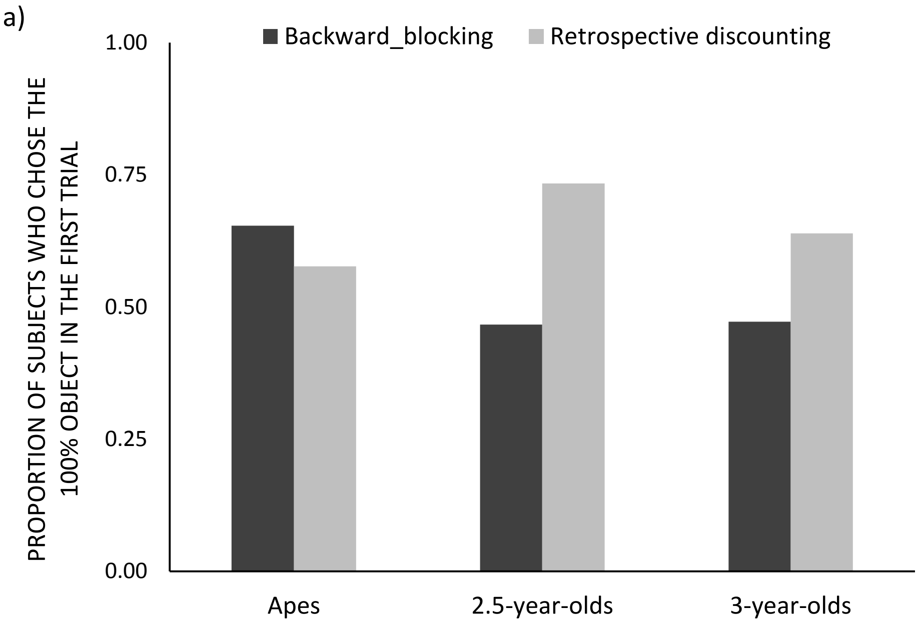
In the test phase, subjects received three different conditions with two trials each: baseline, backward blocking, and retrospective discounting condition (see Table 1). The baseline condition was identical to Experiment 1 (C+, D- or D-, C+). In the backward blocking and retrospective discounting condition, E presented subjects with two novel objects and inserted them together (CD) and the detector was activated. This step was repeated once. In the backward blocking, E inserted object C into the apparatus by itself once, thereby, activating the detector (CD+, CD+, C+). In the retrospective discounting condition, E inserted object D into the apparatus by itself once but the detector did not activate (CD+, CD+, D-). Then the subjects were allowed to choose between object C and D.

Thus, in the retrospective discounting condition the associated effect of C on the detector was conditional on D whereas in the backward blocking condition the effect of C was independent of D. In contrast to Experiment 1, subjects had to re-evaluate the conditional probabilities after the initial, joint CD demonstrations.

3.1.4 Scoring and analysis. Same as in Experiment 1. Overdispersion was not an issue (dispersion parameter: 0.76).

3.2 Results

Figure 3 presents the performance of apes, 2.5-year-olds, and 3-year-olds across condition. A GLMM with group, condition, the interaction of group and condition, and the location of the 100% object was significant compared to the null model (likelihood ratio test: χ^2=24.28, df=8, *p*=0.002). More specifically, we found a significant interaction between group and condition (likelihood ratio test: χ^2=10.02, df=4, *p*=0.040). Pairwise comparisons revealed that 3-year-olds performed significantly better than apes in the retrospective discounting condition (z=2.11, p=0.035) and the baseline condition (z=2.23, p=0.026). Between 3-year-olds and 2.5-year-olds we found no difference in retrospective discounting (z=0.19, p>0.250) or baseline (z=1.11, p>0.250). Similarly, 2.5-year-olds and the apes did not differ from each other in retrospective discounting (z=1.51, p=0.132) or baseline (z=0.87, P>0.250). There were no significant differences between groups in the backward blocking condition (3-year-olds vs. 2.5-year-olds: z=0.55, p>0.250; 3-year-olds vs. apes: z=1.12, p>0.250; 2.5-year-olds vs. apes: z=1.38, p=0.167).



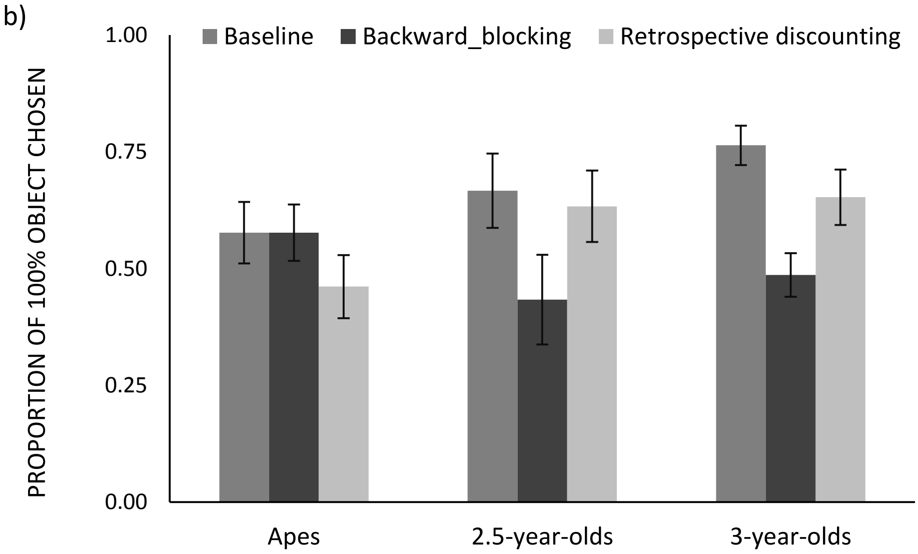


Figure 3. Exp. 2: a) Proportion (and number) of subjects who chose object C in their first trial across conditions; b) Proportion of trials in which object C was selected across the different test conditions (mean ± SE). For each condition, there were two trials per subject. In each trial subjects were presented with a different set of objects. Demonstration: baseline C+, D-; retrospective discounting CD+, CD+, C-; backward blocking CD+, CD+, C+

For 3-year-olds, the condition had a significant effect on performance (χ^2=15.56, df=2, *p*<0.001) with 3-year-olds performing better in the retrospective discounting (z=2.01, p=0.036) and baseline condition (z=3.57, p<0.001) compared to the backward blocking condition. We found no difference between the retrospective discounting condition and the baseline condition (z=1.03, p>0.250). For 2.5-year-olds (χ^2=5.08, df=2, *p*=0.079) and the apes (χ^2=1.92, df=2, p>0.250) we found no significant effect of condition. The location of the 100% object did not have a significant effect on performance (χ^2=2.87, df=1, *p*=0.090).

One-sample t-tests against the hypothetical chance value of 50%, provided no evidence that subjects' choices deviated significantly from chance in the backward blocking condition (apes: t(25)=1.28, p=0.212; 2.5-year-olds: t(14)=0.69, p>0.250; 3-year-olds: t(35)=0.30, p>0.250). In the retrospective discounting condition, 3-year-olds (*t*(35)=2.58, *p*=0.014) but not apes (t(25)=0.57, p>0.250) or 2.5-year-olds (*t*(14)=1.74, *p*=0.104) choose the 100% object significantly more often than expected by chance.

Again, we analyzed whether participants' performance was affected by the trial number within each condition. We found no interaction between trial number and condition for 2.5-year-olds (χ^2=3,92, df=2, *p*=0.141), 3-year-olds (χ^2=0.68, df=2, *p*=0.711), or apes (χ^2=3.00, df=2, *p*=0.224). First trial analysis in session 1 did not reveal significant preferences for object C for apes (backward blocking: 10 of 14 apes, 71.4%, binomial test: p=0.180; retrospective discounting: 8 of 12 apes, 66.7%, p>0.250), 2.5-year-olds (backward blocking: 4 of 8 children, 50%; retrospective discounting: 5 of 7 apes, 71.4%, p>0.250), or 3-year-olds (backward blocking: 8 of 21 children, 38.1%, binomial test: p>0.250; retrospective discounting: 9 of 15 children, 60%, p>0.250).

3.3 Discussion

Three-year-olds, in contrast to the apes, were able to discriminate retrospectively between the two objects based on the retrospective discounting demonstration. The performance of 2.5-year-olds resembled the performance of 3-year-olds even though they did not perform significantly better than expected by chance in the retrospective discounting condition. However, this negative result might be attributed to the small sample size of 2.5-year-olds. In the backward blocking condition, in contrast, neither children nor apes discriminated between the two objects.

Children's performance in the current study is in line with previous research using the blicket detector paradigm. Following backward blocking demonstrations (CD+, CD+, C+), Sobel and colleagues (Sobel & Kirkham, 2006; Sobel, Tenenbaum, & Gopnik, 2004) found that 2- and 3-year-old children, unlike 4-year-olds, did not show a significant preference for object A. In contrast, they showed such a preference in retrospective discounting (CD+, CD+, D-; which they term 'indirect screening-off' condition). Consistent with these findings, Beckers and colleagues (2009) found evidence for retrospective discounting but not for backward blocking in 3-year-olds using a forced-choice procedure like in the current study. McCormack and colleagues (2009) reliably found such retrospective cue competition effects only at the age of 5-6 years. The large variation in the developmental onset of these cue competition effects across studies within the same experimental paradigm has been explained by the involvement of controlled reasoning processes (McCormack et al., 2009). Using different methods such as anticipatory looking paradigms, cue competition effects were found in much younger infants (Sobel & Kirkham, 2006).

What made this task harder for the apes and 2.5-year-olds compared to Experiment 1? The difference compared to Experiment 1 was twofold. First, the current task required retrospective re-evaluation of the evidence whereas in Experiment 1 the apes could discriminate between the two objects based on the first information that they got in the one-cause condition. Second, in experiment 2 we showed the apes whether object C by itself was activating the detector or not. However, in contrast to Experiment 1, we did not show them whether or not object D would activate the detector by itself. Thus, the evidence that the apes received in Experiment 2 was incomplete. Third, apes might have not paid sufficient attention to the experimenter-given demonstrations. To examine whether apes are in fact unable to succeed on such retrospective re-evaluation tasks given sparse evidence we presented them again with these conditions. This time, however, we manipulated whether apes could insert the objects on their own (and observe the effects of their own interventions on the apparatus) or whether they would receive as before demonstrations by a human experimenter.

4 Experiment 3

4.1 Material and Methods

4.1.1 Subjects. Twenty-two chimpanzees, eight bonobos, and seven orangutans participated in this experiment. All subjects except for two juvenile orangutans and three chimpanzees had participated in Experiment 1 and/or 2 before. The interval between the experiment 2 and 3 was on average 157 days (range: 134-169 days).

4.1.2 Materials. The apparatus was the same as in Experiment 1. Sixteen novel objects were used as stimuli.

4.1.3 Procedure and Design. Subjects who did not participate in the test phase of experiment 1 or 2 received an initial training with a single object that activated the apparatus when being inserted by E (see experiment 1). Eight subjects received the initial training. Two of these eight individuals (two juvenile orangutans) started to insert the objects on their own after having received the demonstration and continued with the discrimination training. The six remaining subjects did not insert the objects within four sessions and were excluded from the study.

We administered two test phases: an observation and an intervention phase. We counterbalanced the order of the phases across subjects. The observation phase was identical to the previous experiments, i.e. the subjects received experimenter-given demonstrations before they could choose and insert one of the two objects. E inserted the objects by placing them directly inside the detector via the backside of the apparatus. In the intervention phase, the apes could insert the objects during the demonstration on their own. Here, E handed the objects to the apes according to the demonstration scheme via the central hole in the panel to which the sliding platform was mounted. For compound demonstrations (CD+), we tied the two objects by means of zip ties. Thereby, we ensured that the apes would insert the two objects together in CD+ demonstrations of the intervention condition. After the compound trials, E cut the zip tie in full view of the subject and split the two objects. On a few occasions, apes broke the zip tie. In these trials, we prevented the apes from inserting the separated objects into the apparatus (by blocking access to the tube on top of the detector) and asked the objects back in order to re-tie them. We excluded one orangutan and one chimpanzee from the intervention condition because they managed to insert a single object during a compound trial after they had broken the zip tie.

In contrast to previous experiments, not only the activation of the detector was contingent upon the insertion of certain objects (as before) but also the deactivation of the detector was correlated with the removal of the object from the apparatus. We included this modification to increase the salience of the effect.

In each phase, subjects received one training session and two test sessions. The training consisted of C+, C+, D- , D- demonstrations (in counterbalanced order). Depending on the phase, subjects passively observed experimenter-given demonstrations or could intervene on the apparatus by inserting the objects on their own. Subjects proceeded to the test phase of each condition only if they met the training criterion (same as in the previous experiments: correct choice in every trial at least in their second attempt).

In the test trials of each phase, subjects received the backward blocking (CD+, CD+, C+) and retrospective discounting condition (CD+, CD+, D-). We administered one trial per condition. Subjects received one trial per session. We counterbalanced the order of conditions across subjects.

4.1.4 Scoring and analysis. We used McNemar's test to compare the conditions in each test phase. For the combined analysis of both test phases we used a GLMM. We included condition, test phase, order of test phases, and side of the 100% object in the model as well as subject ID as random effect. We also included random slopes of the side of the 100% object within subject. Overdispersion was not an issue (dispersion parameter: 1.14).

4.2 Results

4.2.1 Training. Twenty-nine individuals inserted the objects during the training phase of the observation and intervention condition (one chimpanzee inserted the objects only in the intervention training and one juvenile orangutan inserted the objects only in the observation training). In both training phases, subjects choose the 100% object significantly more often than expected by chance (intervention: M±SEM 79.3±5.9%, $t(28)=5.03$, $p<0.001$; observation: 65.5±5.4%, $t(28)=2.77$, $p=0.010$). In the intervention phase, 21 out of 29 individuals met the training criterion (six bonobos, 11 chimpanzees, and four orangutans). In the observation phase, 18 out of 29 individuals reached the criterion (three bonobos, 12 chimpanzees, and three orangutans). Thirteen individuals (three bonobos, eight chimpanzees, and two orangutans) passed both trainings.

4.2.2 Test. Figure 4 presents the performance of apes across conditions and test phases.

4.2.2.1 Observation phase. As in experiment 2, we found no significant difference between conditions (exact McNemar's test: p=0.688) and subjects did not score significantly better than expected by chance in any condition (backward blocking: M±SEM 55.6±12.1%, p=0.815, retrospective discounting: 66.7±11.4%, p=0.238).

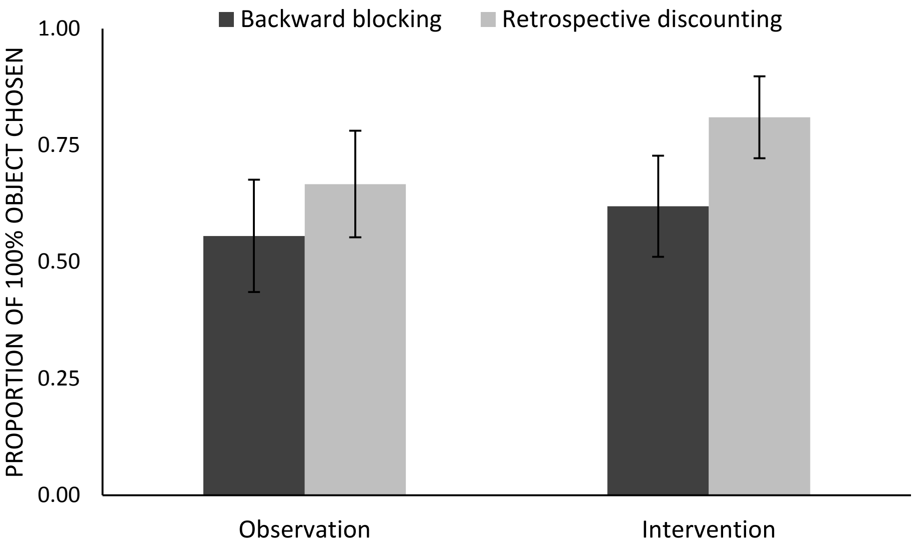


Figure 4. Exp. 3: Proportion of subjects that chose object C across the different test conditions and test phases (mean ± SE). For each condition, there was one trial per subject.

4.2.2.2 Intervention phase. Likewise, we found no significant difference between conditions (exact McNemar's test: p=0.344) when the apes operated the apparatus on their own during the demonstration. However, subjects scored significantly better than expected by chance in the retrospective discounting condition (M±SEM 81.0±8.8%, p=0.007) but not in the backward blocking condition (M±SEM 61.9±10.9%, p=0.383).

4.2.2.3 Comparison between observation and intervention phase. When analyzing the data of the thirteen individuals who completed both test phases, the GLMM did not reveal a significant interaction between observation/intervention and backward blocking/retrospective discounting conditions (χ^2=0.46, df=1, *p*=0.499). After removing the interaction from the model, we found that the apes performed significantly better in the retrospective discounting condition compared to the backward blocking condition (χ^2=4.27, df=1, *p*=0.039). No significant difference was found between the observation and intervention phase (χ^2=0.13, df=1, *p*=0.713), the order of the phases (χ^2=0.19, df=1, *p*=0.666), or the location of the 100% object (χ^2=0.85, df=1, *p*=0.357).

4.3 Discussion

In the current experiment, the apes performed overall significantly better when they received negative information (D-) as compared to positive evidence (C+) about the contingency between one object and the activation of the apparatus. Apes performed particularly well in the retrospective discounting condition of the intervention phase, i.e. when the apes by themselves could intervene on the apparatus. Apes' modest performance in the observation phase replicates the findings obtained in experiment 2: based on observational evidence alone, the apes did not perform above chance in any retrospective re-evaluation condition. Operating at the apparatus by themselves seemed to facilitate the task even though we did not find a significant difference between the observation and intervention phases. Apes' performance in the current experiment resembles the performance of three-year-old children in experiment 2. That is, apes, like three-year-olds, performed overall significantly better in the retrospective discounting condition compared to the backward blocking condition.

The forced-choice procedure employed in the current set of experiments has the notorious problem that the performance may be explained based on the avoidance of one alternative, a preference for the other alternative, or a combination of the two. For the current results this means that apes in the retrospective discounting condition might have merely avoided object D (and chose the only remaining object on the platform, object C) or that they have inferred (by excluding object D as candidate cause) that object C is the most likely cause of the detector onset. The results obtained in the two-cause condition of experiment 1 suggest, however, that a single extinction trial (D-) alone might not necessarily be sufficient to bias apes' choices. Only in combination with compound trials (CD+), single extinction trials had a significant effect on apes' performance (one-cause condition in experiment 1 and retrospective discounting condition in experiment 3) – a finding which is in line with an inference by exclusion account. In an inferential framework, the CD+ demonstration suggests that at least one of the two objects is causing the effect. Without this basis, the D-demonstration does not allow for making any valid inferences about the causal status of object C.

If avoidance alone does not alone explain apes' performance, cue competition effects such as backward blocking and recovery from overshadowing might be at work here. The current findings suggest that apes' performance can be accounted for by recovery from overshadowing but not by backward blocking. From an associative learning perspective, however, both of these cue competition effects are expected under the same circumstances (Beckers et al., 2009). Again, from an inferential point of view, these results are expected because CD+ followed by C+ demonstrations are not informative about the causal status of object D. CD+ followed by D- trials, in contrast, do provide evidence for the causal relationship between object C and the effect.

Notwithstanding the above, methods are desirable that allow by design for disentangling avoidance and active choice explanations. In a follow-up experiment, we presented apes with more complex patterns of demonstration (cf. Beckers et al., 2009) that make differential predictions for the two types of explanation. These demonstrations were based on their own interventions and involved four (instead of two) objects presented as two compound stimuli on the platform. However, the apes failed already during the discrimination suggesting that tracking the increased number of stimuli concurrently located on the platform hampered their performance (see supplementary material).

In our final experiment, we investigated another feature that is often used to differentiate causal from spurious relations, the spatial relationship between cause and effect. Following Gopnik and Sobel (2000), we examined apes' sensitivity to the spatial relationship between the objects and the apparatus during the demonstration.

5 Experiment 4

5.1 Methods

5.1.1 Subjects. Thirteen chimpanzees, eight bonobos, and five orangutans participated in this experiment. All subjects had participated in Experiment 3 before. The interval between the experiment 3 and 4 was on average 77 days (range: 76-85 days).

5.1.2 Materials. The apparatus was the same as in Experiment 1. Sixteen novel objects were used as stimuli.

5.1.3 Procedure and Design. We administered two conditions, the causal and association condition. In both conditions, subjects received experimenter-given A+, A+, B-, B- demonstrations (in counterbalanced order, each trial with a novel set of objects). The causal condition was identical to the previous experiments, i.e. E inserted the objects in the apparatus. In the association condition, the procedure was identical with the exception that E did not insert the objects into the apparatus but held them up next to the apparatus while the apes were sitting at the apparatus and looking towards E. E held the objects up directly at the Plexiglas panel separating E and the apes. For A+ demonstrations, E would then insert her hand into the apparatus and touch the bottom of the chamber (while holding up object A with her other hand). In B- demonstrations E lifted her hand behind the apparatus without inserting it in the apparatus while holding up object B with her other hand. In both conditions, the demonstration lasted for 3 seconds (E controlled the timing by means of a metronome). After these demonstrations, apes were allowed to choose one of the two objects. If they picked object B, the apparatus would not be activated upon insertion of the object and the trial was repeated once.

5.1.4 Scoring and analysis.

Same as in Experiment 1. We included condition, order of conditions, session, side of the 100% object, and species in the model as well as subject ID as random effect. We also included random slopes of condition, order of condition, session, and side of the 100% object within subject. Overdispersion appeared to be no issue (dispersion parameter: 0.84).

5.2 Results

We found that the order of conditions had a marginally significant effect on apes' performance (see Figure 5). Apes tended to perform better in their second condition compared to the first one (χ^2=3.77, df=1, *p*=0.052). In the association condition, apes performed significantly better when they received the association condition after the causal condition as compared to when they received the association condition first (χ^2=3.86, df=1, *p*=0.049). In the causal condition the order did not have a significant effect on performance

(χ^2=1.19, df=1, p=0.274). Apart from this order effect, the GLMM did not reveal any other significant effect (condition: χ^2=1.35, df=1, p=0.245; species: χ^2=1.40, df=2, p=0.496; session: χ^2=2.54, df=1, p=0.111; side of the 100% object: χ^2=0.06, df=1, p=0.811).

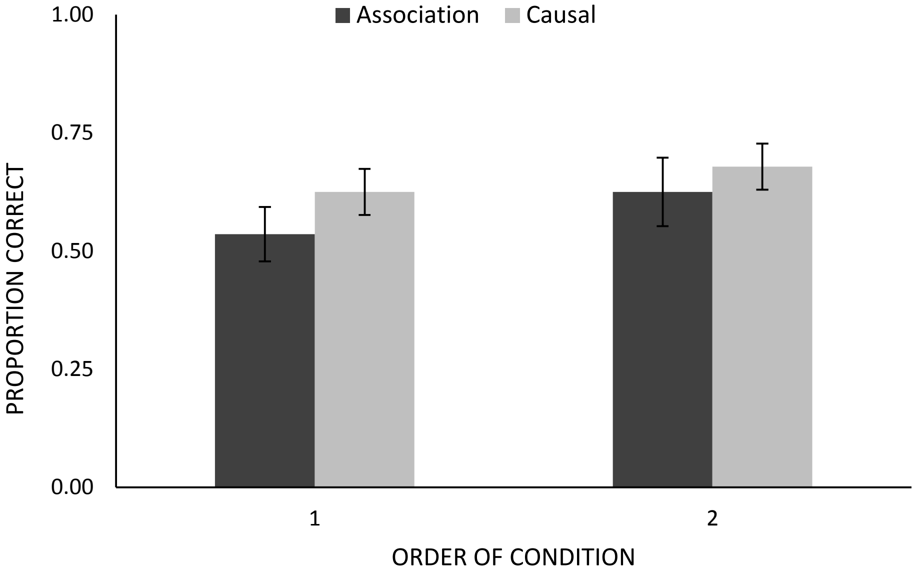


Figure 5. Exp. 4: Proportion of trials in which object A was selected across the different test conditions and order of administration (mean ± SE). For each condition, there were four trials per subject.

Apes selected the 100% object significantly above chance levels in the causal condition (t(25)=4.50, p<0.001) but not in the association condition (t(25)=1.69, p=0.103). Irrespective of the order in which the conditions were administered, apes performed significantly above chance in the causal condition (1^{st}: t(11)=2.57, p=0.026; 2^{nd}: t(13)=3.68, p=0.003) but their performance did not deviate significantly from chance in the association condition (1^{st}: t(13)=0.62, p=0.547; 2^{nd}: t(11)=1.73, p=0.111).

After errors had occurred (i.e. when apes chose object B), we repeated the trial and let them choose again. In the association condition, apes tended to repeat the error. They showed a significant preference for the incorrect object (i.e. the object that they had already chosen before) in their second attempt (t(22)=-2.23, p=0.036). In the causal condition, apes' second choice did not deviate significantly from the hypothetical chance level of 50% (t(22)=-0.49, p=0.628).

5.3 Discussion

The apes chose the 100% object significantly above chance levels only when the object made contact with the apparatus (causal condition) but not when the object was only held up by the experimenter next to the apparatus (association condition). Moreover, in the association condition apes were less flexible when they made errors and tended to repeat incorrect choices - something that was not the case in the causal condition. We note that the effects found in this experiment were rather weak and partially driven by an order effect (apes performed better in the association condition after having received the causal condition before). Nevertheless, the results suggest that apes' performance in the causal condition was more robust than in the association condition.

The causal and association condition shared some important procedural features: in both conditions, the experimenter was lifting up one of two objects, which was followed (or not) by the onset of the apparatus. In the causal condition this effect was produced upon the insertion of the object whereas in the association condition the hand of the experimenter was inserted in the apparatus while the object was held up next to the apparatus. Apes' above-chance performance in the causal condition suggests that the insertion of the object into the apparatus facilitated the task for them.

It is conceivable that the apes have learned in the course of the previous experiments to pay attention in particular to what happened inside the apparatus. We made sure that apes would look at the objects during the demonstration in both conditions and that the time the apes would encounter the contingency between objects and the state of the apparatus would be identical. Nevertheless, the apes might have directed their attention more to what happened inside the apparatus than what happened next to it. Whether or not physical contact between object and apparatus was crucial for apes' choices in this task can therefore not conclusively determined here. At the very least, our results suggest that apes learned to pay attention to the spatial relationship between the objects and the apparatus and not merely to the actions of the experimenter (e.g. lifting one of the objects) and its contingency with the food reward.

6 General discussion

Our findings suggest that apes, like young children, take patterns of variation and covariation into account to infer novel causal structures under some circumstances. Based on this information they produce their own interventions to bring about the desired effect. More specifically, apes and 2.5-year-olds demonstrated sensitivity to conditional independence in a completely novel and arbitrary context (experiment 1). Importantly, their choices were based on the minimal number of observations required for this kind of cue competition effects. That is to say, the crucial piece of information (the A+ and B- demonstration) was only presented once and the participants encountered each set of objects only once. Nevertheless, the apes and 2.5-year-olds discounted the confounded information already in the first trial of the first test session. Furthermore, the mostly negative results of experiment 2 indicate that using conditional independence to make retrospective causal judgments might be more difficult for the apes and 2.5-year-olds. Only by the age of 3 years, we found solid evidence that children were able to update their causal judgments retrospectively. Experiment 3 suggests, however, that apes do succeed on retrospective discounting (but not backward blocking) tasks when they observe the effects of their own interventions. The pattern of findings of experiment 1 and 3 suggests that a simple avoidance strategy was insufficient to account for apes' performance. Finally, experiment 4 showed that apes' choices and error corrections were modulated by the spatial relationship between the apparatus and the objects.

The fact that apes and children were able to differentiate between completely novel objects based on patterns of conditional independence after a single demonstration shows that they learn novel causal structures very efficiently. A purely associative explanation for the current findings on apes' and children's sensitivity to conditional independence are cue competition effects such as forward blocking (Kamin, 1969) and protection from overshadowing (Blaisdell, Bristol, Gunther, & Miller, 1998). Kamin found that the conditioning of an association between a conditioned stimulus (CS, e.g. tone) and an unconditioned stimulus (US, e.g. food) is inhibited by the presence of a second CS (e.g. a light) which has already been associated with the US before, a phenomenon known as forward blocking. Accordingly, in the one-cause condition the participants might have learnt by classic conditioning during the (single) A+ demonstration to expect food when object A is inserted. In the two subsequent AB+ demonstrations, the established association between A and the activation of the detector might have "blocked" the conditioning of the association between object B and the activation of the detector. In addition, it is possible that the B-demonstration might have led to latent inhibition of object B which, in turn, protected object A from overshadowing effects during AB+ demonstrations (Beckers et al., 2009). Both of these cue competition effects might have yielded stronger associations of object A (relative to object B) with the activation of the detector in the one-cause condition as compared to the two-cause condition. However, forward blocking or protection from overshadowing effects by latent inhibition are usually based on multiple exposures to the precondition trials (A+ or B-) CS in nonhuman animals (e.g. in the study by Blaisdell et al., 1998, rats received 120 latent inhibition presentations, B-, before they were presented with the overshadowing trials, AB+). There are some reports on one-trial blocking in rats (e.g., Azorlosa & Cicala, 1986; Dickinson, Nicholas, & Mackintosh, 1983; Gillan & Domjan, 1977). However, “one-trial” refers here to a single presentation of compound trials (AB+) but multiple preconditioning trials (A+). Moreover, this type of one-trial blocking is found in specific contexts such as taste aversion and fear conditioning. The current findings differ markedly from these instances: “blocking” occurs after only a single exposure to the preconditioning stimulus (A+) in a foraging context and the apes produced their own interventions on the apparatus on the basis of the observed patterns of evidence. Likewise, the fact that 3-year-olds and apes succeed in the retrospective discounting (or recovery from overshadowing) condition but not in the backward blocking condition is not expected on the basis of associative models of cue competition effects.

Therefore, the current findings raise the possibility that these associative cue competition processes work on different time scales in great apes as compared to rats for instance. However, simple stimulus-outcome association tasks using, for example, token-exchange paradigms (in which subjects need to learn that tokens with certain perceptual features, unlike others, are rewarded) provided no evidence that nonhuman apes are characterized by particularly fast associative learning rates (e.g. Hanus & Call, 2011; Pelé, Dufour, Thierry, & Call, 2009; Schrauf & Call, 2009; see also Call, 2006). Alternatively, apes’ performance in the current task might be based on qualitatively different, inferential reasoning processes.

In human children, the results of the blicket detector paradigm have been explained by the integration of patterns of correlations into directed causal maps (Gopnik et al., 2004). A major advantage of such a representation is that it can explain sensitivity to conditional independence and the ability to predict the effects of one’s own interventions. In the context of the blicket detector paradigm, Gopnik and colleagues (2001) have shown in a follow-up experiment that 3- and 4-year-old children were also able to use information about conditional probabilities to invent a novel intervention on the detector. Future studies might elucidate whether apes, like human children, are also able to craft novel interventions in order to deliberately resolve ambiguous causal structures. Similar to the findings with rats (Blaisdell, Sawa, Leising, & Waldmann, 2006) and children (e.g., Schulz, Gopnik, & Glymour, 2007), it will be an interesting question for future research whether apes might also use their interventions to distinguish between more complex causal structures such as common cause and causal chain models.

Irrespective of the precise learning algorithm, apes and children preferred evidence generated by unconfounded interventions (i.e., manipulating a single variable) rather than confounded interventions (i.e., manipulating multiple variables simultaneously). Only the former provide evidence for an unconditional dependence between the manipulated variable and the observed effect. Unconfounded interventions are, therefore, crucial for inferring causal structures (Woodward, 2011). Woodward differentiates between different sources of information for making causal judgments: own interventions, other agents' interventions, and observed covariation. All three sources of information, however, may be interpreted in a counterfactual way; that is, as what would happen if someone was intervening on the task at hand. In the current study, apes efficiently learned causal structures based on own interventions supporting Penn and Povinelli's notion (2007) that great apes explicitly represent causal relations between own actions and their effects. However, apes also learned causal relations from observing other agents' interventions. Whether apes, like humans, also distinguish between conditional and unconditional dependencies between events based on observed covariation information alone (i.e. without any agent involved, cf. the wind example from the beginning) is unclear to date and should be addressed by future research.

Our findings might be particularly relevant for the acquisition of tool-use in great apes. Causal learning might be the cognitive ability that sets tool-users apart from non-tool-users. In contrast to recent findings with New Caledonian crows (Taylor et al., 2014) , great apes produced causal interventions based on sparse observational evidence. However, as mentioned before it is still unclear whether the poor performance of NC crows is related to this particular task or is indicative of a deeper limitation in producing and learning from causal interventions (Jacobs et al., 2015). Future studies should be aimed at systematically comparing tool-using and non-tool-using species on causal learning and intervention tasks such as the blicket detector paradigm.

Finally, the current findings also link with the literature on social learning in great apes. Nonhuman apes seem to learn from others predominately by means of emulation (i.e. copying environmental results, end states, or goals) rather than by imitating others' actions (e.g., Call, Carpenter, & Tomasello, 2005; Horner & Whiten, 2005; Tennie, Call, & Tomasello, 2006; Tennie, Call, & Tomasello, 2010; Tomasello, Davis-Dasilva, & Camak, 1987). For emulation learning, identifying causally relevant interventions is crucial as it allows individuals to reproduce an effect in the environment without copying the exact actions of a model. Imitation, in contrast, works even without any knowledge about the task-relevant contingencies. Our results support the emulation account by demonstrating how apes identify causally relevant interventions when they observe another agent's actions by discounting confounded ones.

In summary, the current study provides evidence that apes, like preschoolers, are able to make accurate causal judgments based on sparse observational evidence of patterns of variation and covariation under some circumstances. Candidate explanations for apes' performance are cue competition effects as well as inferential reasoning processes. Apes and children made these causal judgments after only a single exposure to the task-relevant contingencies, which makes associative cue competition effects (such as blocking) appear less likely. Instead, inferential reasoning processes might be at work here. However, this tentative conclusion awaits further confirmation by future research.

Acknowledgements

C.J.V. was supported by a scholarship of the German National Academic Foundation. We thank Raik Pieszek and Henrik Röthel for constructing the apparatus, Maike Foraita for her help with reliability coding, and the animal caretakers of the zoo Leipzig.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Azorlosa, J. L., & Cicala, G. A. (1986). Blocking of conditioned suppression with 1 or 10 compound trials. *Animal Learning & Behavior, 14*(2), 163-167.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 238-249.

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General, 135*(1), 92-102.

Beckers, T., Vandorpe, S., Debeys, I., & De Houwer, J. (2009). Three-year-olds' retrospective revaluation in the blicket detector task: Backward blocking or recovery from overshadowing? *Experimental psychology, 56*(1), 27-32.

Bird, C. D., & Emery, N. J. (2009). Insightful problem solving and creative tool modification by captive nontool-using rooks. *Proceedings of the National Academy of Sciences, 106*(25), 10071-10072.

Biro, D., Inoue-Nakamura, N., Tonooka, R., Yamakoshi, G., Sousa, C., & Matsuzawa, T. (2003). Cultural innovation and transmission of tool use in wild chimpanzees: evidence from field experiments. *Animal Cognition*, *6*, 213-223.

Blaisdell, A. P., Bristol, A. S., Gunther, L. M., & Miller, R. R. (1998). Overshadowing and latent inhibition counteract each other: Support for the comparator hypothesis. *Journal of Experimental Psychology: Animal Behavior Processes, 24*(3), 335-351.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*(5763), 1020-1022.

Call, J. (2006). Descartes' two errors: reasoning and reflection from a comparative perspective. In S. Hurley & M. Nudds (Eds.), *Rational animals* (pp. 219-234). Oxford: Oxford University Press.

Call, J., Carpenter, M., & Tomasello, M. (2005). Copying results and copying actions in the process of social learning: chimpanzees (*Pan troglodytes*) and human children (*Homo sapiens*). *Animal Cognition, in press*.

De Houwer, J., Vandorpe, S., & Beckers, T. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior, 33*(2), 239-249.

Dickinson, A., Nicholas, D., & Mackintosh, N. (1983). A re-examination of one-trial blocking in conditioned suppression. *The Quarterly Journal of Experimental Psychology, 35*(1), 67-79.

Dobson, A. J. (2002). *An introduction to generalized linear models*. Boca Raton: Chapman & Hall/CRC press.

Edwards, B. J., Rottman, B. M., & Santos, L. R. (2011). The Evolutionary Origins of Causal Cognition. In T. McCormack & C. Hoerl & S. Butterfill (Eds.), *Tool Use and Causal Cognition* (pp. 111-128). Oxford: Oxford University Press.

Edwards, B. J., Rottman, B. M., Shankar, M., Betzler, R., Chituc, V., Rodriguez, R., Silva, L., Wibecan, L., Widness, J., & Santos, L. R. (2014). Do Capuchin Monkeys (*Cebus* *apella*) Diagnose Causal Relations in the Absence of a Direct Reward? *PloS one, 9*(2), e88595.

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology, 65*(1), 47-55.

Gillan, D. J., & Domjan, M. (1977). Taste-aversion conditioning with expected versus unexpected drug treatment. *Journal of Experimental Psychology: Animal Behavior Processes*, *3*(4), 297-309.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review, 111*(1), 3-32.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*(5), 1205-1222.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620-629.

Hanus, D., & Call, J. (2011). Chimpanzee problem-solving: contrasting the use of causal and arbitrary cues. *Animal Cognition, 14*(6), 871-878.

Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British journal for the philosophy of science, 50*(4), 521-583.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science, 317*(5843), 1360-1366.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Animal Cognition, 8*(3), 164-181.

Hume, D. (1748 / 2000). *An enquiry concerning human understanding: A critical edition* (Vol. 3): Oxford University Press.

Jacobs, I., von Bayern, A., Martin-Ordas, G., Rat-Fischer, L., & Osvath, M. (2015). *Corvids create novel causal interventions after all.* Paper presented at the Proc. R. Soc. B.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York: Appleton-Century-Crofts.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science, 16*(9), 678-683.

Lefebvre, L., Reader, S. M., & Sol, D. (2004). Brains, innovations and evolution in birds and primates. *Brain, behavior and evolution, 63*(4), 233-246.

Manrique, H. M., & Call, J. (2011). Spontaneous use of tools as straws in great apes. *Animal Cognition, 14*(2), 1-14.

Manrique, H. M., Gross, A. N. M., & Call, J. (2010). Great apes select tools on the basis of their rigidity. *Journal of Experimental Psychology: Animal Behavior Processes, 36*(4), 409-422.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009). Cue competition effects and young children’s causal and counterfactual inferences. *Developmental psychology, 45*(6), 1563-1575.

Nazzi, T., & Gopnik, A. (2000). A shift in children’s use of perceptual and causal cues to categorization. *Developmental Science, 3*(4), 389-396.

Pelé, M., Dufour, V., Thierry, B., & Call, J. (2009). Token transfers among great apes (*Gorilla gorilla*, *Pongo pygmaeus, Pan paniscus*, and *Pan troglodytes*): species differences, gestural requests, and reciprocal exchange. *Journal of Comparative Psychology, 123*(4), 375-384.

Penn, D., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*(2), 109-129.

Penn, D., & Povinelli, D. (2007). Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology, 58,* 97-118.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology, 20*(2), 416-420.

Schrauf, C., & Call, J. (2009). Great apes' performance in discriminating weight and achromatic color. *Animal cognition*, *12*(4), 567-574.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science, 10*(3), 322-332.

Seed, A. M., & Call, J. (2009). Causal knowledge for events and objects in animals. In S. Watanabe & A. P. Blaisdell & L. Huber & A. Young (Eds.), *Rational animals, irrational humans* (pp. 173–187). Tokyo: Keio University Press.

Seed, A. M., Hanus, D., & Call, J. (2011). Causal knowledge in corvids, primates, and children. In T. McCormack & C. Hoerl & S. Butterfill (Eds.), *Tool Use and Causal Cognition* (pp. 89-110). Oxford: Oxford University Press.

Sloman, S. A. (2009). *Causal models: How people think about the world and its alternatives*. New York, NY: Oxford University Press.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental psychology, 42*(6), 1103.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*(3), 303-333.

Taylor, A. H., Cheke, L. G., Waismeyer, A., Meltzoff, A. N., Miller, R., Gopnik, A., Clayton, N. S., & Gray, R. D. (2014). Of babies and birds: complex tool behaviours are not sufficient for the evolution of the ability to create a novel causal intervention. *Proceedings of the Royal Society of London B: Biological Sciences, 281*(1787), 20140837.

Tennie, C., Call, J., & Tomasello, M. (2006). Push or pull: Imitation vs. emulation in great apes and human children. *Ethology, 112*(12), 1159-1169.

Tennie, C., Call, J., & Tomasello, M. (2010). Evidence for emulation in chimpanzees in social settings using the floating peanut task. *PloS one, 5*(5), e10544.

Tomasello, M., & Call, J. (1997). *Primate cognition*. New York, NY: Oxford University Press.

Tomasello, M., Davis-Dasilva, M., & Camak, L. (1987). Observational learning of tool use by young chimpanzees. *Human Evolution, 2*(2), 175-183.

Völter, C. J., & Call, J. (2014). Great apes (*Pan paniscus, Pan troglodytes, Gorilla gorilla, Pongo abelii*) follow visual trails to locate hidden food. *Journal of Comparative Psychology, 128*(2), 199-208.

von Bayern, A. M. P., Heathcote, R. J. P., Rutz, C., & Kacelnik, A. (2009). The role of experience in problem solving and innovative tool use in crows. *Current Biology, 19*(22), 1965-1968.

Woodward, J. (2011). A Philosopher Looks at Tool Use and Causal Understanding. In T. McCormack & C. Hoerl & S. Butterfill (Eds.), *Tool Use and Causal Cognition*. Oxford, UK: Oxford University Press.