BRILL *Behaviour (2018)* DOI:10.1163/1568539X-00003528 **Behaviour** brill.com/beh

Logical reasoning by a Grey parrot? A case study of the disjunctive syllogism

Irene M. Pepperberg [a,*], **Suzanne L. Gray** [a,b], **Shilpa Mody** [a,c],
Francesca M. Cornero [a] **and Susan Carey** [a]
[a] Department of Psychology, Wm James Hall, Harvard University,
Cambridge, MA 02138, USA
[b] Department of Psychology, Tufts University, Medford, MA 02155, USA
[c] Current address: Facebook, 1 Hacker Way, Menlo Park, CA 94025, USA
[*]Corresponding author's e-mail address: impepper@wjh.harvard.edu

Received 2 September 2018; initial decision 26 September 2018; revised 1 November 2018;
accepted 8 November 2018

Abstract

In Call's (2004) 2-cups task, widely used to explore logical and causal reasoning across species and early human development, a reward is hidden in one of two cups, one is shown to be empty, and successful subjects search for the reward in the other cup. Infants as young as 17-months and some individuals of almost all species tested succeed. Success may reflect logical, propositional thought and working through a disjunctive syllogism (A or B; not A, therefore B). It may also reflect appreciation of the modal concepts "necessity" and "possibility", and the epistemic concept "certainty". Mody & Carey's (2016) results on 2-year-old children with 3- and 4-cups versions of this task converge with studies on apes in undermining this rich interpretation of success. In the 3-cups version, one reward is hidden in a single cup, another in one of two other cups, and the participant is given one choice, thereby tracking the ability to distinguish a certain from an uncertain outcome. In the 4-cups procedure, a reward is hidden in one cup of each pair (e.g., A, C); one cup (e.g., B) is then shown to be empty. Successful subjects should conclude that the reward is 100% likely in A, only 50% likely in either C or D, and accordingly choose A, thereby demonstrating modal and logical concepts in addition to epistemic ones. Children 2 1/2 years of age fail the 4-cups task, and apes fail related tasks tapping the same constructs. Here we tested a Grey parrot (*Psittacus erithacus*), Griffin, on the 3- and 4-cups procedures. Griffin succeeded on both tasks, outperforming even 5-year-old children. Controls ruled out that his success on the 4-cups task was due to a learned associative strategy of choosing the cup next to the demonstrated empty one. These data show that both the 3- and 4-cups tasks do not require representational abilities unique to humans. We discuss the competences on which these tasks are likely to draw, and what it is about parrots, or Griffin in particular, that explains his better performance than either great apes or linguistically competent preschool children on these and conceptually related tasks.

DOI 10.1163/1568539X-00003528

Keywords
parrot cognition, inference by exclusion, disjunctive syllogism, Grey parrot.

1. Introduction

Since Descartes (1637/1985) at least, philosophers have debated whether the capacity for abstract, combinatorial, propositional thought is unique to humans. Philosophers, historically, and modern cognitive scientists, currently, have argued for both sides of the debate (human uniqueness: e.g., Descartes, 1637; Davidson, 1982; Penn et al., 2008; versus widely attested in the animal world: e.g., Fodor, 1975; Tomasello & Call, 1997; Griffin, 2001; ten Cate & Healy, 2017). Philosophers' arguments are a priori, but the issue is not one to be settled from the armchair. It is a straightforward scientific matter, which involves both a theoretical analysis of what kind of thought is actually being investigated and the design of appropriate empirical studies that bring data to bear on whether such thought is manifest in nonhuman animals. Empirical research on these issues necessarily proceeds case study by case study.

One case study taken to show abstract, combinatorial, propositional thought concerns deductive inferences involving logical connectives — specifically, the disjunctive syllogism (A or B; not A, therefore B). A behaviour that shows an animal working through a disjunctive syllogism transparently reveals such thought.

In pursuit of evidence that nonhumans can work through a disjunctive syllogism, subjects from a range of species have been tested on what is called in the animal cognition literature "reasoning by exclusion", which we will call "the process of elimination" (P of E). In an early test of P of E reasoning, Premacks' cup task (Premack & Premack, 1994), two containers are baited with two different foods of equal desirability while the subject watches. One food is taken out of one of the containers such that the animal cannot see which one, and the food is shown to the animal (e.g., by the experimenter eating it or putting it in a pocket). The subject is then supposed to choose the cup from which food had not been removed to obtain its reward. Many researchers claim that this behaviour is consistent with working through a disjunctive syllogism: Food is in container A or B; it is not in A; therefore it is in B (Premack & Premack, 1994; see also Call, 2006). Some of Premack's chimpanzees (*Pan troglodytes*), Call's apes (*Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo pygmaeus*), one of seven tested Clark's nutcrackers

(*Nucifraga columbiana*, Tornick & Gibson, 2013), and a number of Grey parrots (*Psittacus erithacus*, Mikolasch et al., 2011; Pepperberg et al., 2013) succeeded on this task.

Of course, researchers are aware that a more conservative interpretation of these animals' behaviour is possible in this particular task. For instance, the subjects could simply have learned that the presence of the exhibited food is a cue to avoid the cup that contained that food. The Grey parrot studies included controls for this alternative interpretation of success. After baiting, nothing was removed from the cups, but one type of food was shown to the birds, by taking it out of a pocket and replacing it. The parrots then responded by choosing between the cups at random, demonstrating that merely the sight of a particular food did not cue their behaviour. In contrast, on the actual P of E test trials, the Grey parrots clearly inferred which cup was empty from seeing the identity of the removed food, and then chose the other cup. In one Grey parrot study (Pepperberg et al., 2013), another series of trials also showed that birds were not merely avoiding any cup from which food was withdrawn.

A more widely used version of the task was developed by Call (2004), possibly based on a task first used by Grether & Maslow (1937). In Call's task, animals see that one of two hidden containers is baited with food, whereupon the containers are revealed, and one is shown to be empty (see also Schmitt & Fischer, 2009; Hill et al., 2011). Success consists in searching in the remaining cup for the food. Spontaneous success has been observed in a wide variety of nonhuman species (see Voelter & Call, 2017, Table 2). In an initial study with human children, Hill and colleagues (Hill et al., 2012) found that 3–5-year-old children succeed at the Call (2004) cup task. Subsequent work has found spontaneous success at 23 months, 20 months and 17 months (but not 15 months: Mody & Carey, 2016; Feiman et al., 2017; Mody et al., data not shown).

Nonetheless, the interpretation of successful P of E behaviours as reflecting a deductive inference based on the disjunctive syllogism may not be warranted in any of these experiments. There is no evidence that the inference leads to deductive certainty, and some evidence suggests that it actually does not do so. We know of four studies that have addressed the issue of deductive certainty, involving either nonhuman animals or young children, or both. Watson and colleagues (Watson et al., 2001) compared dogs and 5-year-old children on an invisible displacement task in which the subjects watched an experimenter walk behind three equidistant barriers with a toy, and emerge from the last with empty hands (in fact, he had not deposited the toy behind any of the three spatially separated barriers). The experimental question was whether, for a subject that had searched behind two of the barriers, the time to switch to the third would be faster than the time to switch from the first to the second, due to the inference that the toy must be behind that remaining barrier. This pattern was observed for the 5-year-old humans, but the opposite was observed for the dogs, whose behaviour slowed, as if the motivation to search was being extinguished. Thus, 5-year-olds showed a signature of a deductive inference but the dogs did not.

Call & Carpenter (2001) carried out an experiment with a logic closely related to that of the Watson et al. dog experiment. Chimpanzees and 2 1/2-year-old human children were presented a task consisting of trials in which one of three tubes contained a treat. They could lean over and look into each tube, but this required some effort. It was worth doing so because it was not easy to extract the treat from the tube. Call and Carpenter's main goal was to establish whether apes and children monitored if they had or had not seen the treat placed into a tube, looked into the tubes if they had not yet seen the location in which it was hidden, and terminated their search and chose a tube when they had seen that it held the treat. That is, the experiment concerned monitoring knowledge versus ignorance, where the behaviour of "seeking information" was the metric that reflected a state of ignorance. Both populations provided evidence of selective information seeking as a function of having seen or not seen the hiding of the treat. Both populations were highly likely to choose a tube if they'd seen the treat placed in it, or had seen the treat by looking and finding that the treat was there, and the children were much more likely to look into one or more tubes if they were ignorant of the treat's location. However, the apes chose without looking into any of the tubes on more than half of the trials in which they were ignorant of the location and even the children did so on about a quarter of the trials.

Relevant to the issue of deductive inference, Call and Carpenter reported what they called "hyper-efficient" searches, searches in which the participants chose the third tube without looking after having looked in two tubes and finding each empty. Apes did this on 13.9% of the trials in which they had looked in two empty tubes; children did so on 4.6% of such trials. Two of 10 apes did this at least once; and 2 of 8 children did this at least once. Call and Carpenter present this as evidence of working through a disjunctive syllogism — of inferring where the object is upon learning where it is not — but many aspects of their data counter this interpretation. First, this behaviour is actually extremely rare, in both populations. Second, Call and Carpenter also present data on the frequency of "insufficient searches" — the proportion of times when participants choose a tube without searching after having looked in only one of the tubes and finding it empty. These proportions do not differ from the hyper-efficient searches (12.5% for apes, 3% for children). In sum, while providing very interesting data concerning the relations between seeing/not seeing and seeking information (the main goal of the study), the above results can be seen as a striking demonstration of a failure of deductive inference in both populations, extending Watson et al.'s (2001) failure from dogs to chimpanzees, and demonstrating failure in children at age 2 1/2, in contrast to success at age 5.

Mody & Carey (2016) adopted a different approach to explore whether, for young children, success on this task actually reflects working through a disjunctive syllogism. They assumed toddlers, like nonhuman animals, are not merely avoiding an empty cup, but pointed out that success is consistent with two quite different reasoning schemes (see Figure 1). In one scheme (left half of figure), the disjunctive syllogism, the possible locations of the treat are represented as true disjuncts, joined by an "or", such that learning that the treat is not in one, guarantees that it is in the other. In the other (right half), each location is represented as a possibility, such that eliminating one of the possibilities leaves the other one open as a possibility. The key difference between the two reasoning schemas is that learning that one location is empty leads to an updating of the probability that it is in the other (to 100%) in Figure 1e, but in Figure 1f leads to *no change* in this probability (but keeps it more than 0). It is likely impossible to establish whether an animal or a young child has assigned probability 100% to the remaining option, but Mody & Carey (2016) devised a task that establishes whether learning that one location is empty at least leads to an updating of the probability that the treat is in the other location. This is a minimal condition on P of E reasoning actually reflecting the subject's working through a disjunctive syllogism.

The key to Mody & Carey's (2016) method (see Figure 2) was to show children two sets of cups, hiding one sticker in each of two pairs of cups. Upon being shown one cup is empty, the child working through a disjunctive syllogism should choose the other cup in that pair, whereas the child merely

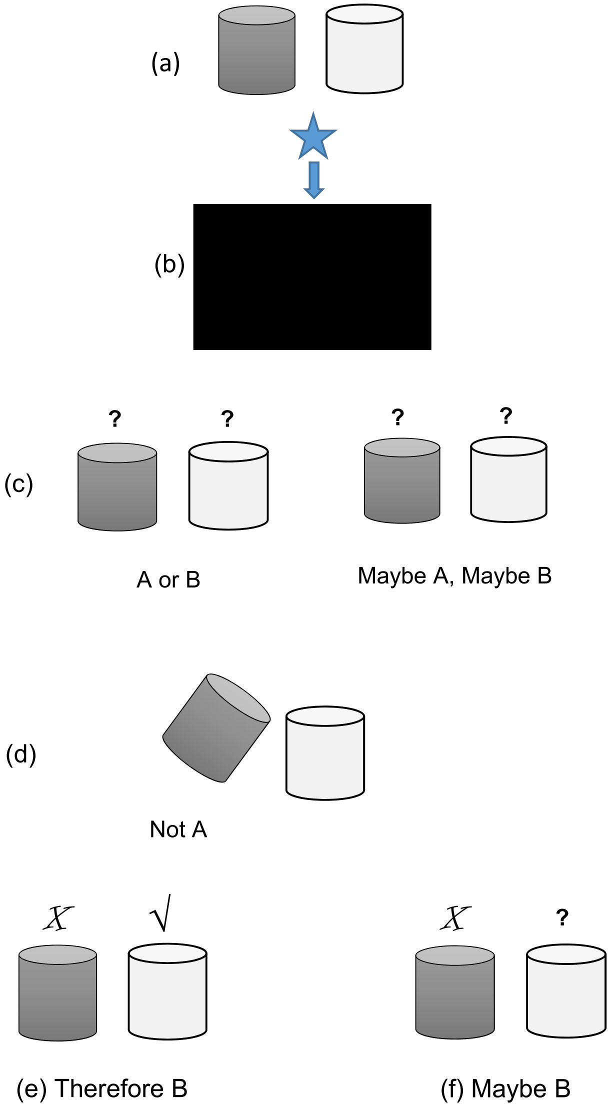

Figure 1. (a) Two distinguishable cups are presented to the subject. (b) A screen is placed between the subject and the cups, and the experimenter shows that a treat is being placed behind the screen, ostensibly in one of the cups. (c) The subject can have two hypotheses as to where the treat has been hidden — that it must be in either A or B, or that it is possibly in A or possibly in B. (d) The subject is shown that the treat is definitely not in A. (e) The subject uses the information and the A or B hypothesis to deduce "not A, therefore B". (f) The subject uses the information and the "maybe A, maybe B" hypothesis simply to cross A off from the possibilities, leaving "maybe B" as an alternative.

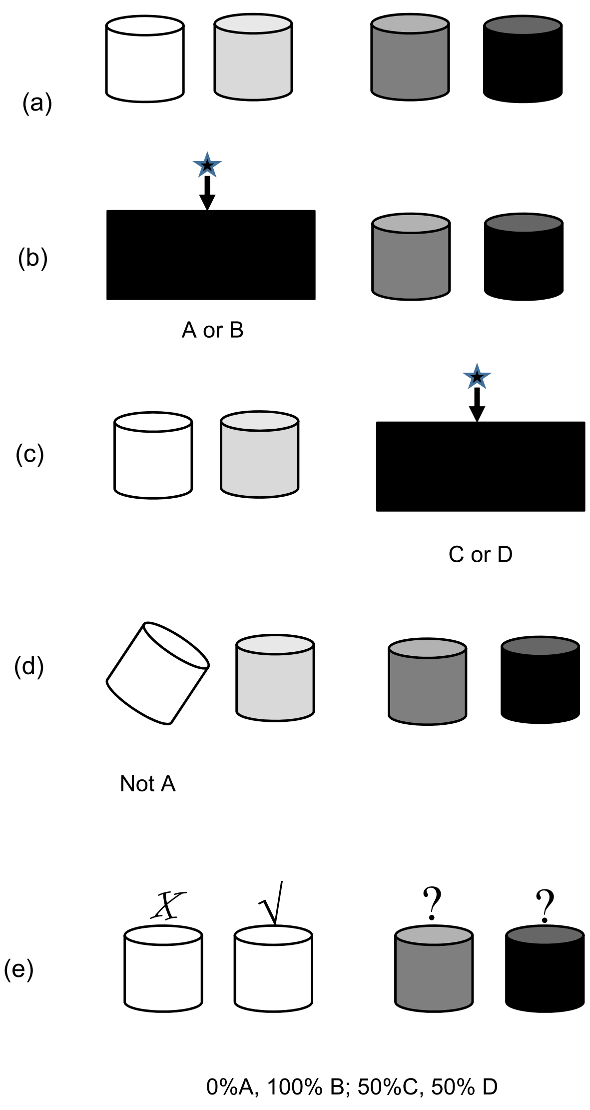

Figure 2. (a) Four distinguishable cups are presented to the subject. (b) A screen is placed between the subjects and the cups on one side, and the experimenter shows that a treat is being placed behind the screen, ostensibly in one of the cups. (c) The same actions are performed on the other side. (d) The subject is shown that the treat is definitely not in A. (e) If the subject understands a disjunctive syllogism, the reasoning should be that there is 100% chance of the reward being in A and only a 50%–50% chance for C, D. A subject that does not understand a disjunctive syllogism will still reason “maybe B, maybe C, maybe D” and choose at chance among those cups. Adapted from Mody & Carey (2016).

eliminating a possibility should look equally in the three remaining cups, because they all are still possible locations of stickers. Mody and Carey showed that 2 1/2-year-olds displayed the latter pattern of choices, searching in each of the remaining locations 36% of the time (no different from chance). Mody (2016) generalized this developmental pattern to P of E in causal reasoning, the "indirect screening off" trials in a blicket detector task. Children are told that they must find the "blicket", an object that activates a novel device such that it will exhibit an interesting event. They were then shown that one pair of blocks (A and B), placed on a device together, activates the device, and then that another pair of blocks (C and D), placed on the device also did so. They then were shown that one of the blocks (A), in isolation, failed to do so, and were given all 4 blocks and asked to "make the device go" by finding the blicket trigger. At age 2 1/2, children chose equally among the remaining blocks (B, C and D). The data from the 4-cups and 4 possible blickets tasks add to the evidence from Call & Carpenter (2001) that 2 1/2-year-olds fail to show behaviour that would signify their working through a deductive inference.

In Mody's studies, 3–5-year-olds performed above chance on both the 4-cups task and the 4 possible blickets task. These successes were consistent with working through the disjunctive syllogism, rather than merely eliminating the option of the empty cup or the inert block without updating the probability of the other option that was linked by *or* to the eliminated option. However, success on the 4-cups task may still be consistent with non-propositional reasoning strategies, for example, probabilistic reasoning over mental models not encoded propositionally (Rescorla, 2009; see 8. General discussion below). Nonetheless, failure on this task provides strong evidence that success on the standard Call 2-cups task is <u>not</u> based on propositional reasoning formulated using the concepts *or* and *not*. In the remainder of this article, whenever the concepts *or* and *not* are mentioned, italics will be used.

Mody and Carey noted that there are two quite different possible interpretations of the failure of 2 1/2-year-old children in the 4-cups and 4-blickets experiments. First, learning the language of disjunction may play an important role in success on these tasks, and 2 1/2-year-olds have not yet done so. Studies of language production reveal use of the word "or" at 3:0 but not before (Bloom et al., 1980; Lust & Mervis, 1980; French & Nelson, 1985; Morris, 2008). It is possible either that learning the meaning of the word "or" is part of the process of acquiring the logical concept *or*, or that learning the meaning of the word "or" makes the concept *or* more available for encoding the disjunctive relation between the hidden stickers and the cups, and the blocks and activation of the blicket machine. Second, the 4-cups task places heavy working memory demands on the child, and preschool children are greatly deficient, relative to adults, in their capacity for updating models of hidden objects held in working memory (e.g., Gathercole et al., 2007). Control tasks provided some evidence that problems with updating working memory cannot entirely explain the difficulty of the 4-cups and 4-blickets P of E tasks for young children, but the question is still open as to whether the majority of the developmental improvement on the 4-cups task between ages 2 1/2 and 4 may be a consequence of an increase in the ability to update working memory.

A recent study of working memory capacity of Griffin, a Grey parrot who successfully completed an earlier 2-cups experiment (Pepperberg et al., 2013), shows that further examination of his abilities provides an ideal opportunity to decide between the hypothesis that knowledge of the linguistic logical connective "or" is necessary for the signature deductive inference probed in the 4-cups case and the hypothesis that performance limitations due to poor working memory may have masked conceptual competence in 2 1/2-year-olds. Notably, Griffin, although trained to use English speech to identify objects, colours, shapes, and to make some requests, knows no words for any logical connectives (not "not, or, and, if..."). However, Griffin's visual working memory, including his capacity for mental updating of working memory models, is almost identical to human adults and vastly superior to that of 7-year-old humans (Pepperberg & Pailian, 2017), and as noted above, preschool children have markedly less working memory capacity than do 7-year-olds. Griffin also already had experience with, and had demonstrated competence in, tasks that reflect other conceptual abilities needed for success on the 4-cups task. Besides logical reasoning, both the 4-cups task and a related 3-cups version (see below) require the representation of probabilities (distinguishing a probability of 1 from probabilities of 0.5), and an understanding that the tasks require choice of the certain option. Griffin has previously been shown to base choices on representations of probability, although the task that was used did not require representations of certainty. Specifically, when required to vocally predict the identity of a sample based on information about its ratio in a given population, he performed at a level comparable to 6- and 7-year-old children on the Piagetian version of this task (Clements et al., 2018). Preschool children, in contrast, respond randomly on the very same task (Piaget & Inhelder, 1975).

Given Griffin's demonstrated competence on important prerequisites of the 4-cups task, his results on this task would provide important cross-species comparisons. His failure would be an important addition to the failures to find signatures of deductive reasoning in nonhuman animals and very young children. It would also provide strong evidence that success on this task is not guaranteed by robust updating of working memory, demonstrated capacity for basing judgments on representations of probability, and demonstrated capacity for P of E reasoning on the 2-cups task. In contrast, Griffin's success on the 4-cups task would place him at the level of at least 5-year old children, would conclusively demonstrate that mastery of the language of logical connectives is not a task prerequisite, and would be consistent with the conclusion that parrots are capable of deductive inferences over propositional representations involving the logical concepts *or* and *not*.

2. General methods

2.1. Subject and housing

Griffin, a male Grey parrot (*Psittacus erithacus*), 20 years old at the start of the experiment, had been the subject of cognitive and communicative studies since his acquisition from a breeder at 7.5 weeks of age. Housing and day-to-day care outside of sleeping conditions are described in Pepperberg & Wilkes (2004); conditions were maintained after moving to Harvard in July, 2013. Food and water were always available. This study used four of his documented abilities: to understand vocal commands such as "wait" (Koepke et al., 2015), to successfully demonstrate full Stage 6 object permanence (Pepperberg et al., 1997), to succeed on simple tests of exclusion (Pepperberg & Wilcox, 2000; Pepperberg et al., 2013), and to track probabilities (Clements et al., 2018).

2.2. Materials

During testing, Griffin sat on the same T-stand used for several other experiments (e.g., Pepperberg et al., 2013). The testing apparatus consisted of a flat, white laminate square tray upon which opaque, equally-sized silver metal cups were placed. An erect cardboard divider, of roughly the same height as the cups, separated the tray into two sections and the cups into appropriate groupings. The tray was lightly marked with a semicircle for the cup locations so that each cup would be equidistant from the parrot's beak when he was centred on the perch, and consistently positioned in each trial. For the 3-cups exclusion task, two cups were placed on one side of the divider and one cup in the other. The placement of the cups (one versus two) on either side was randomized via Random.org. For the 4-cups exclusion task, two cups were always placed on either side. For both tasks, each cup had a circular white cardboard cover, upon which a coloured woollen pompon was attached to differentiate each cup: colours used were red, green, yellow, and blue, so that cups were distinguished by colour as in the Mody & Carey study (2016). Red and green covered cups always appeared to Griffin's right side of the tray, and yellow and blue always appeared on his left, either together or separately according to the number of cups specified by the trial. A large cardboard rectangle screened the cups from Griffin's view while treats were being hid. The tray was placed on a stool at the same height as the T-stand. Raw cashew halves were used to reward correct responses; in one experiment, Skittles© were also used as a reward. Two experimenters sat next to each other in front of the apparatus and across from the parrot during testing (Figure 3); both were involved in hiding the rewards. One experimenter (IMP) was always to the parrot's left; one of four different undergraduate students (depending upon the task) was always to the parrot's right. In 4-cups trials, one (IMP) demonstrated the empty cup; on all trials both noted Griffin's choices and, when necessary (see below) agreed as to whether to call a mistrial and abort the trial. In addition, from Experiment 2 on, all trials were videotaped for later scoring with respect to intercoder reliability. The videotapes were later coded with respect to the choices Griffin made by a research assistant not involved in these studies; coder reliability between the coder and the on-line observers was 100%. To ensure that Griffin was not cued as to which side to search, both experimenters looked only at Griffin during the choice period (after the cups were revealed on the 3-cups trials, and after the four cups had all been touched on the 4-cups trials). Notably, parrots do not respond to human gaze direction when objects on a tray are closely spaced and the human head is about 30 cm away from the tray, as in this study (Giret et al., 2009; Koepke et al., 2015).

Figure 3. Apparatus set up for Griffin before the beginning of each 4-cups trial.

2.3. Noncompliance procedures/mistrials

As had been the case in other studies with Griffin (e.g., Pepperberg & Nakayama, 2016) and another Grey parrot, Alex (e.g., Pepperberg & Carey, 2012), Griffin occasionally engaged in noncompliant behaviour. These instances were considered mistrials and were not included in our data set. Specific conditions had to be met for Griffin's behaviour to be considered noncompliant and a mistrial declared: Griffin would have to fall asleep on or tumble from the perch, be distracted by the behaviour of the other bird in the laboratory (as evidenced by his looking towards that bird instead of the experimental set-up), be engaged in a preening bout from which he could not be distracted, or repeatedly request to return to his cage (vocalizations of "Wanna go back"). Additionally, in this study, experimenters could cause a mistrial by erring in their performance. If a mistrial occurred, Griffin received a 5-min break in testing. If a second mistrial followed a first, the session was aborted. When mistrials occurred, the trials were placed at the end of the experiment and redone. Sessions were series of trials planned for a single day, usually four trials long, as described in the methods for each experiment.

The classification of mistrials was made with the agreement of both experimenters on-line. Mistrials were rare; the number of trials repeated because of mistrials, and number of sessions that were aborted in each experiment is noted in the results sections. For Experiments 2–5, which were videotaped, videotaping occurred up to the point that Griffin completed the trial or the experimenters agreed to declare a mistrial (i.e., with no information as to what choice Griffin had made).

2.4. Training procedures

Griffin's experience with these tasks as we formalized the procedure is described fully in the Appendix. Because Griffin was already familiar with the P of E from the previous 2-cup exclusion studies (Pepperberg et al., 2013), we began with two attempts to test him on the 4-cups task. He failed at both of these: one due to a colour bias explainable in hindsight because of intervening experiments in which one particular colour (green) had been reinforced (Péron et al., 2013, 2014), and one because he had injured a foot and could not easily choose cups on one side of the display (leading to side biases). We allowed the foot to heal, and decided to test him on the 3-cups procedure (see below, reported as Experiment 1, which Mody and Carey used as a training procedure as well as control trials for the 4-cups task. That is, the 3-cups task constitutes training for and practice with aspects of the 4-cups task that are not present in the 2-cups tasks on which he had previously been tested — namely, distinguishing certainty from mere possibility; distinguishing probability of 1 from probability of 0.5. It does not, however, provide training relevant to the P of E inference.

As detailed in the Appendix, after Experiment 1 he failed at another attempt at the 4-cups procedure, which reflected re-injury of his foot, which was confirmed by his now failing at the 3-cups procedure of Experiment 1. We then let his foot heal for a full year before carrying out Experiments 2–5. In sum, the 3-cups task is a form of training relevant to the 4-cups task, and Griffin also had some 60 trials of experience over two years with the 4-cups task where he showed side biases (due to injured foot) or the colour biases noted above. In these 60 trials, he received feedback from his choices that food is always found in the cup from the same pair as the empty cup, and is only found half of the time if a cup from the other pair is chosen. This feedback did not lead to improved performance over trials when he was showing a colour bias, or a side bias due to his injured foot. Moreover, he received no training on the final version of the 4-cups task.

3. Experiment 1: the 3-cups task

In the Autumn of 2014, after the procedural issues due to cup colour were resolved (see Appendix), we tested Griffin on the 3-cups task that was used as training trials for children (see Mody & Carey, 2016). This task is not a test of exclusion, because certainty is guaranteed by the fact that a nut is always lowered behind the screen on the side of the divider with one cup. Success, however, would demonstrate that Griffin was able to follow the two hiding events, was motivated to get a reward on his first choice, would confirm he could distinguish probabilities of 1 from probabilities of 0.5, and would choose the cup where he could be certain to obtain the reward. Very young children, including even the 2 1/2-year-olds who choose randomly among the three potentially non-empty options on the 4-cups task, perform better than chance on the 3-cups task (i.e., assuming equal probability of choosing any cup). However, performance by preschool children is far from ceiling. Only around half of the choices at ages 2 1/2 and 3 are of the certain cup. Although this performance (47 and 60%, respectively) is robustly better than chance (33% choice of this cup), it is clear that this task is very hard for preschoolers: Given that distinguishing where they would have the higher likelihood of success involves choosing the side with one cup over the side with two cups, these children are at or only slightly above chance (two sides, 50% probability). Even by ages 4 and 5, the certain cup is chosen only 71 and 72% of the time, respectively (Mody & Carey, 2016).

The 3-cups task is interesting in its own right. Hanus & Call (2014) allowed chimpanzees to make a choice of a single container among two sets of containers, in which the probability of finding food in any given container differed between the sets. They found that chimpanzees reliably chose a container from the set with a higher probability of a reward, but only when the ratios of probabilities in the two sets was greater than 2:1. The 3-cups task has this structure, and the ratio is 2:1, a ratio at which the chimpanzees chose randomly between the two sets. In addition, in the Hanus and Call procedure, the choice was strictly based on the ratios of the probability; there was no special status of probabilities of 1 (certainty). Granted, the Hanus and Call task differed from the present task in how it created different probabilities of rewards in the cups in the two sets. If, however, the present task also elicits a choice based on the ratio of probabilities of rewards, and Griffin, like apes, could neither distinguish probabilities unless that ratio is greater than 2:1 nor distinguish certainty from mere possibility or uncertainty, he should fail the 3-cups task.

3.1. Method

Trials began with the placement of three cups on the white tray. A small cardboard barrier separated the pair and the singleton cup. Cups were left uncovered, with the covers behind each cup; Griffin was shown and verbally told that each cup was empty ("Look... empty!"). The tray was moved slightly away, and a cardboard screen was placed between Griffin and the cups, obscuring his view. One experimenter showed Griffin half a cashew, then moved the nut behind the screen and hid it in one of the cups on that side of the tray. The other experimenter repeated the actions, of showing half a nut and hiding it, on her side of the tray. The order in which the experimenters acted and the cup in which the cashew was placed on the 2-cup side were randomized via Random.org. With the blind still in place, the experimenters covered the cups. They then removed the blind, told Griffin to "wait", and then pushed the tray slightly towards Griffin, indicating that he could choose a cup ("Go find the nut"). Both experimenters vocalized equally. Griffin would lean forward and uncover one cup with his beak. If the cup contained a nut, he was allowed to eat it and received verbal praise (e.g., "Good boy"). Again, both experimenters vocalized equally. If the cup was empty, the tray was pulled back so that Griffin could not uncover another cup, and he was told he was wrong (e.g., "Sorry, no nut"). Forty trials were conducted in this manner, generally with four but occasionally with up to five trials occurring in the same session. A session was ended before four trials were completed if Griffin exhibited noncompliant behaviour, and extended to five if he refused to leave the T-stand after the fourth trial. Whether or not he found the nut, his response was considered "correct" only if he chose the cup that was alone on one side of the tray.

3.2. Results

In the first session, Griffin either fell asleep or did not attend to the task during each of the four trials (all marked as mistrials; the session was not aborted until the fourth trial as it was his first session). Starting with the second session, he was correct on 33/40 consecutive trials (82.5%), erring on trial numbers 13, 14, 20, 21, 29, 34 and 40; on a binomial test, chance of 0.33, $p \ll 0.001$ (see Figure 4). He erred no more than once in any single session. No additional mistrials occurred.

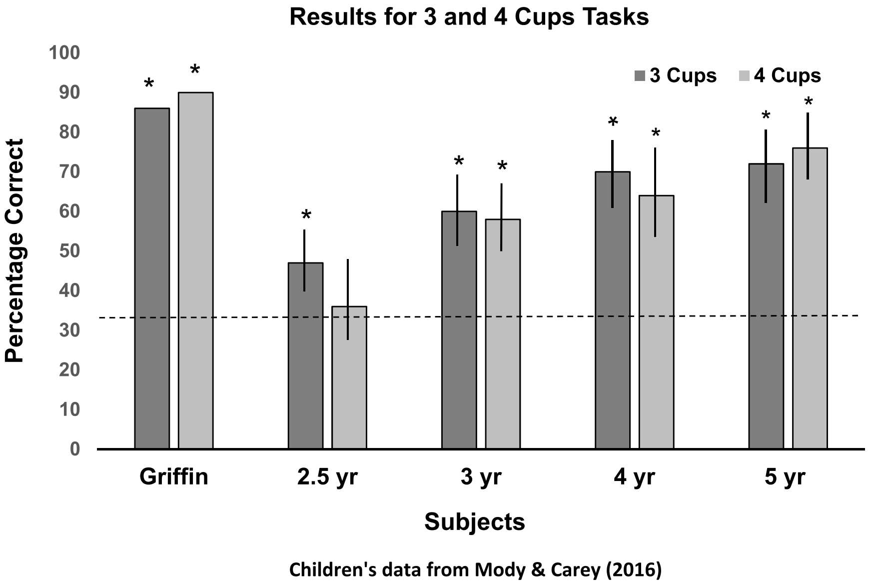

Figure 4. Percentage of correct responses on the 3-cups and standard 4-cups trials for Griffin (all trials on all experiments combined) compared to children in the Mody & Carey (2016) study. All but the 2 1/2-year-old children on the 4-cups tasks are statistically significant, $p \leqslant 0.01$.

3.3. Discussion

This task does not involve reasoning by exclusion, but rather probes the subject’s understanding of the difference between a certain payoff and a 50/50 chance of a payoff as well as his motivation to get the food on his first choice. Both are prerequisites for success on the 4-cups task, which in addition requires a P of E inference to arrive at a representation that the food is certainly in a specific cup. Interestingly, by making the certain choice approx. 83% of the time, he outperformed 5-year-old children on this task, who choose the correct cup only 72% of the time, and greatly outperformed 3-year-olds and 2 1/2-year-olds, who are correct only 60 and 47% of the time, respectively. The data from the 2 1/2-year-olds are in line with those from the apes in the Hanus & Call (2014) study, who chose the certain cup 44% of the time for those trials with exactly the same structure (i.e., a choice between a certain cup — one reward in one cup — versus a pair of uncertain cups — one reward in one of two cups). The similarity extends further. In a separate study with 3-year-olds, Andreuccioli (2018) found that children chose the certain cup (1 sticker in one cup) 52% of the time when pitted against a 0.5 probability (1 sticker in one of two cups) and 79% of the time when pitted against a 0.25 probability (1 sticker in one of four cups). The comparable results for Hanus & Call (2014) were, 44 and 78%, respectively.

We return to the question of whether Griffin is distinguishing certainty from mere possibility in 8. General discussion.

4. Experiment 2: the basic 4-cups task

Our plan to proceed immediately was delayed because of an injury to Griffin's foot in the Autumn of 2015 (see Appendix). We therefore waited until he was completely healed, until Summer, 2016.

4.1. Method

At the start of a trial, cups were arranged with the colour and number per side specified as described in 2. General methods. Cups were left uncovered, with the covers behind each cup; as in Experiment 1, Griffin was shown and verbally told that each cup was empty. The tray was moved slightly away, and the cardboard screen was placed between Griffin and the cups, obscuring his view. This time, each experimenter simultaneously held one half-cashew and showed it to Griffin over the blind, informing him they each had a nut, and then hid the cashews in the pre-determined cups (established via Random.org) behind the blind. With the blind still in place, the experimenters covered the cups. They then removed the blind and told Griffin to "wait". The cover of an empty cup from one side (again determined via Random.org) was removed, and the experimenter indicated both vocally and by showing him that it was empty. The cup was then covered again, all four cups were briefly touched by one experimenter so that the last cup touched would not be privileged, and the tray was pushed toward Griffin, indicating that he could choose a cup ("Go find the nut"). If he chose the cup with the cashew on the side on which he had been shown the empty cup, his choice was considered correct and he was allowed to eat the cashew while receiving verbal praise. If he chose the cup that he had been shown to be empty, or any cup on the other side, his response was considered incorrect, but he was allowed to eat the cashew if he grabbed it before the experimenters could react. Sixteen trials were conducted in this manner, with up to four trials occurring in the same session; he could be given up to three sessions per week. A session was ended early if Griffin showed known signs of noncompliance or the experimenters were unable to be present for a total of four trials.

4.2. Results

On the basic 4-cup exclusion task, Griffin was correct 15/16 trials (94%; Figure 4). There were no mistrials or aborted sessions. He hesitated in his choice on only two trials (i.e., did not move forward immediately after being shown an empty cup), on one of which he then proceeded to choose correctly and on one of which he then erred, actually choosing the demonstrated empty cup. His performance is well above a conservative estimate of chance (33%, assuming that he would avoid the demonstrated empty cup and therefore leaving only three remaining options), $p < 0.01$ (binomial). Although he may have learned something about the task while we were formalizing the procedure (see Appendix), he did not exhibit learning over the test trials of Experiment 2: He was correct on his first seven trials; his one error was on his eighth.

4.3. Discussion

These data show that, as for 3–5-year old children, there was no cost to Griffin for arriving at certainty from an inference (94% correct) relative to the 3-cups training/control trials (83%). Furthermore, Griffin's success on the 4-cups task was more categorical than for any age group of human children tested by Mody and Carey (2 1/2-year-olds through 5-year-olds, percent choice of the certain cup 36 to 76%, respectively; Griffin, 94%; Figure 4). However, it is possible that over several iterations of this task as we formalized the procedure, he could have learned an associative rule that would have generated success: choose the cup next to the cup that had been shown to be empty. Of course, this rule cannot explain his success on the 3-cups task (Experiment 1). Experiments 3 and 4 sought to replicate his success on the 4-cups task, as well as test whether the associative alternative is a viable account of that success.

5. Experiment 3: 4-cups task replication with type-1 gambling trials

One way to tease apart his use of a simple association versus an inferential process is to introduce trials in which use of the former would lead to failure. One way of doing so would be to hide a nut in only one of the two pairs of cups, and then to show him that one of the cups from the side with no nut was empty. If his success on the 4-cups trials derives solely from a learned, associative, strategy, “choose the cup next to the one shown to be empty”, he should continue to search in the remaining cup from the same side as that shown to be empty in both the standard 4-cups trials and these new trials in Experiment 3. In contrast, if he were using an inferential process, the latter trials should induce him to choose randomly between the two cups on the side with a nut.

5.1. Method

In this series, in the Autumn of 2016, 16 of the trials were repeats of the standard 4-cups exclusion but 8 additional trials were “gambling” trials. In gambling trials, we replicated the format of the standard 4-cups exclusion trials, but showed him that we were hiding only a single nut on one side. The only difference between the standard trials and the gambling trials is that one experimenter showed that her hand was empty before both experimenters lowered their hands behind the screens. We then removed the blind, and showed Griffin an empty cup on the side that did not contain any nut. The order of trials, placement of the nut, insertion of gambling trials, and the choice of cups were determined by Random.org. If Griffin was reasoning from a representation of the contents of each set of cups, he would know that the cup next to the empty one now was not a correct choice; he should then avoid following what could have been the simple association formed earlier, and should gamble on one of the cups on the other side to try to gain a reward.

5.2. Results

There were three mistrials, which were redone at the end of the other trials. In these trials he was either distracted by another bird in the lab when we were performing the trial ($N = 2$) or we erred by showing him the wrong cup ($N = 1$). Griffin did not hesitate on any of the scored trials, immediately making a choice when presented with the tray.

On the standard 4-cups (non-gambling) trials included in Experiment 3, he was correct on $13/16$ (81%) (binomial test, chance 0.25), $p < 0.01$. Because he picked the demonstrated empty cup on all errors (trials 9, 19, 23), we assumed he considered all four cups a viable option. But even on the more conservative estimate of 33% chance choice of the correct cup, this behaviour is significantly better than chance ($p < 0.01$). He appropriately gambled on 6/8 of the gambling trials (75%), winning a nut by correct guesses on four trials and stealing a nut on a fifth before we could pull the tray away. On the only two trials in which he refused to gamble, he chose the demonstrated empty cup. The choice of the correct cup on the 4-cups trials (81%) was significantly more likely than the choice of the cup next to the demonstrated empty cup on the gambling trials (0%); Fisher's exact test (two-tailed), $p < 0.01$.

5.3. Discussion

These data provide strong evidence against the hypothesis that his success on the standard 4-cups trials was due to the associative strategy "pick the cup next to the demonstrated empty cup". Rather, his three errors on 4-cups trials were choices of the demonstrated empty cup itself (as was one of his two errors in Experiment 2), as were his two deviations from gambling on gambling trials. This behaviour suggests he experienced some difficulty inhibiting attention to the manipulated cup, which was perhaps exacerbated by frustration related to the uncertainty inherent in the gambling trials.

One might argue that success on the gambling trials reflected merely his avoiding the empty side. This is, of course, the rational action. That this may have been the source of his behaviour does not undermine the conclusion that his overall performance cannot be explained by a learned strategy of choosing the cup next to the cup shown to be empty. To provide a further test of the associative strategy, we designed another replication of the standard 4-cups trials intermixed with a different set of gambling trials. Experiment 4 tests the hypothesis that Griffin would gamble even if treats were on both sides of the display. Experiment 4 also avoids the possible frustration due to being faced with trials in which there was no certainty of reward.

6. Experiment 4: 4-cups task replication with type-2 gambling trials

Experiment 4 was carried out in the Winter of 2016–2017. We gave Griffin a choice between a sure-bet cashew and a 50–50 chance of something he would like even better, and for which he therefore might be willing to gamble. We used a Skittle©, one of Griffin's most favoured treats. These candies are not healthy if provided on a regular basis, so their availability was a very special event. Successful gambling in this task could not be due to the strategy of avoiding a side in which no reward had ever been placed, and would confirm that Griffin was not simply following a learned strategy of choosing the cup next to one shown to be empty.

6.1. Method

Experiment 4 had the same overall design as Experiment 3, with 16 regular 4-cups trials involving cashews and 8 gambling trials, except that on these gambling trials we used a Skittle©. Now Griffin saw us hide a nut on one side and a candy on the other; we showed him an empty cup on the side with the nut. If he inferred that the remaining cup from the set that includes the empty cup contains a nut, he knew he had a sure choice of a nut or a 50% chance of a more highly-valued reward. We expected that, at least on some of these trials, he would be willing to gamble, and thus not simply go to the cup next to the one that had been shown to be empty. To ensure that he would be encouraged to gamble here after his failures to get a nut on three of the gambling trials in Experiment 3, we rigged the first gambling trial, surreptitiously placing candies under both cups and guaranteeing at least one win. Because he stopped gambling after a loss, we also surreptitiously rigged his sixth gambling trial: In case he decided to gamble again after having picked the sure-bet nut in several trials, we could learn if a success would again encourage gambling. Note that we had no real expectation that he *would* gamble on the sixth trial, but *we* gambled that enough time had passed since a loss that he would be willing to test his luck again. He was not forced nor encouraged to gamble on this trial. He was not being trained to gamble, as only two of eight widely separated trials were rigged, and we had no particular reason to expect him to gamble on either of those — or actually on any — of his trials; choosing a sure bet over a gamble would be a perfectly rational behaviour.

6.2. Results

There were four mistrials. In three mistrials, he was either not attending or falling asleep; in the fourth, the cup cover shook loose as the experimenter removed the blind and he could see the nut. These four trials were repeated at the end of the experiment.

Again, he robustly succeeded at the standard 4-cups trials. He chose the certain cup on 15/16 trials (94%; binomial test, chance 0.33, $p \ll 0.01$). His one error (trial 14) was to choose a cup from the other side (and lose).

On the gambling trials he went for the Skittle© on 5/8 trials (62.5%), winning a candy on only three trials. He won a candy after gambling on the rigged first trial; he then gambled on the next such trial and lost. He then refused to gamble on the next three gambling trials after this loss. In those instances, he preferred to choose the sure-bet nut. Fortuitously for us, he did gamble again on the sixth gambling trial, which we had rigged. After winning he was again willing to gamble. He won on the next and lost on the last gambling trial.

Again, his behaviour on gambling trials differed markedly from his behaviour on the standard 4-cups trials. He chose the cup next to the empty cup on 94% of the 4-cups trials and on 38% of the gambling trials (the difference is statistically significant, Fisher's exact test, two-tailed, $p < 0.01$). These data rule out the hypothesis that his success on 4-cups trials solely reflects a learned association between rewards and the cup next to the empty cup. Unlike his actions in previous experiments, he never chose the demonstrated empty cup in this experiment. His pattern of behaviour suggests that he was representing the probabilities accurately, sometimes choosing the sure cashew on gambling trials and sometimes taking a chance to get the more highly valued Skittle©, while practically always choosing the sure bet on the standard 4-cups trials interleaved with the gambling trials (see Figure 5).

Is it possible that by ensuring his gamble paid off on two of the gambling trials we taught him to gamble? This is unlikely — he gambled on the first trial he could, whereas he never chose the side opposite to the empty cup on the first trial of any of the three replications of the standard 4-cups trials. Two reinforced choices are certainly insufficient to establish a learned strategy. Too, if we look at his choices before he received a reward on that 6th trial, he chose to gamble on three of the first six gambling trials; thus, even if we disregard the trials after the reward on the 6th trial, the difference between choosing a cup next to the empty cup on standard versus gambling trials at that point is still statistically significant (Fisher's exact test, two-tailed, $p < 0.05$). Furthermore, over all standard 3- and 4-cups trials he chose the uncertain side eight times, and in two of these choices found the nut by chance. These trials in which the gamble on the uncertain side led to a nut did *not* lead to an increase of choices of the uncertain side in the 3- and 4-cups standard conditions.

6.3. Discussion

When given the choice to gamble for a bonus treat or choose the sure-bet nut — in the latter case to choose the cup next to the empty — he sometimes

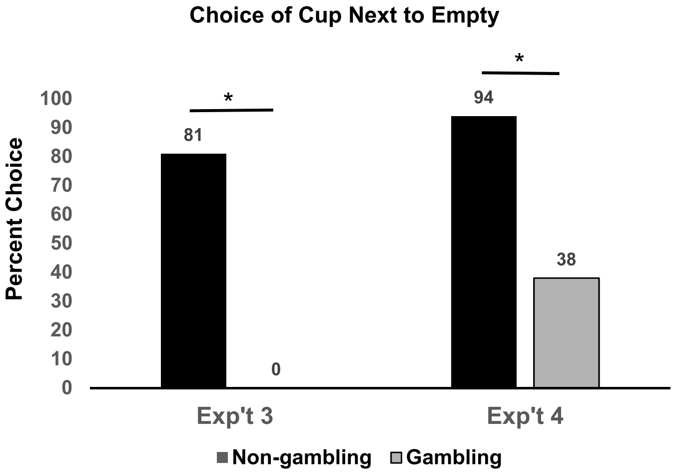

Figure 5. Percentage of trials in which Griffin chose the cup next to the empty cup on the gambling and non-gambling trials in Experiments 3 and 4. The differences were statistically significantly different in both experiments, $p \leqslant 0.01$.

did choose to gamble. Interestingly, he was rather conservative, in that his decision to gamble was sometimes influenced by his prior success or failure. After receiving the bonus treat, he was willing to take the next immediate risk, but refused to gamble for a while after a failure. Again, the insertion of gambling trials did not affect his behaviour on the standard 4-cups task. He was not following a single simple rule, but rather basing his decisions on inference by exclusion and evaluation of risk.

7. Experiment 5: 3-cups task replication

As detailed in the Appendix, Griffin failed the 3-cups procedure during one of the two times he had an injured foot. The above experiments, after his foot had healed, include three successive replications of the standard 4-cups experiment. The last study, carried out in Spring, 2017, was an exact replication of the 3-cups experiment, Experiment 1, from two and a half years earlier. If his successes on the 4-cups task involve tracking probabilities, distinguishing probabilities of 1 from those of 0.5, and/or distinguishing certainty from uncertainty, rather than some combination of simpler rules (avoid the cup next to the empty one, avoid the empty side, choose the side with a Skittle©, etc.), he should certainly continue to track probabilities and succeed.

7.1. Results and discussion

There were no mistrials. On the replicated 3-cups exclusion task, Griffin was correct on 20/20 trials (100%), $p \ll 0.001$. He never hesitated, immediately going to the single cup as soon as he was presented with his choices. He clearly was not guessing, and the gambling trials had not affected his understanding of the task. Across all of these studies, there is strong evidence that Griffin was systematically choosing the cup certain to contain a nut over a cup in which it was merely possible there was a nut, that he would gamble on a merely possible option if it was impossible that the other choice had a treat (Experiment 3), and would sometimes do so if the possible option had a much more highly desirable treat than the certain option (Experiment 4).

8. General discussion

This study establishes that success on the Mody & Carey (2016) 4-cups task does not require linguistic symbols — e.g., “or” or “not” — for the logical connectives, and that it is achievable by at least one nonhuman animal. Over three replications of 16 trials each, Griffin, a Grey parrot who does not know linguistic expressions for any logical connectives, chose the certain cup roughly 90% of the time. Like human children, ages 3 to 5, his performance on the 3-cups trials in which arriving at certainty does not require an inference, and on the 4-cups trials, in which it does, did not differ. His performance on both types of trials was equivalent to that of the 5-year-old children tested in Mody & Carey (2016), even exceeding their score on the percentage correct.

These results raise several related questions. Most importantly, what is the nature of the representations and computations that underlie Griffin’s success on this task? Does Griffin’s near ceiling performance on this task establish that parrots have the capacity for deductive reasoning involving concepts such as *or* and *not*? Also, what is it about Griffin that makes him better at these tasks (both the 3- and 4-cups versions) than young preschool children, and relatedly, what is it about Griffin that yields a signature of deductive reasoning in the face of failures to find such a signature in great apes? We emphasize the importance of these comparisons of his abilities to those of young children and great apes, given that over 300 million years of evolution separate parrots from humans (Hedges et al., 1996). But another reason to be interested in these comparisons is that additional related experiments, discussed below, on both apes and children, provide further insight into what is limiting their performance, limits that Griffin apparently does not have.

8.1. What representations/computations underlie Griffin's success?

We can confidently conclude that at the very least Griffin has the following representational capacities. First, he makes mental models of locations of the treats within cups (their possible locations) that incorporate appropriate probabilities. Second, in the 3-cups task, he assigns a higher value to the probability that a treat is in the certain cup than to the probabilities it is in any of the uncertain cups. Furthermore, in the 4-cups task, he must rule out a possibility upon being shown that a given cup is empty, and then restrict his updating of the remaining probabilities to the other cup in the same set, raising it, at the very least, to a higher probability than that in either of the uncertain cups in the other set. The data are certainly compatible with the conclusion that Griffin creates non-linguistic representations of propositions, including the logical connectives *or* and *not*. Importantly, these experiments also potentially draw upon additional logical capacities as well — modal concepts such as "possible", "necessary", "impossible". Griffin distinguishes certainty, that may derive from deductive necessity, from mere probability, in his consistent — and almost perfect — choice of the certain location over the location with only a 50% chance of containing a treat, in both the 3- and 4-cups tasks. He also rejects the impossible outcome in Experiment 3. Representations of possibilities are at least in some circumstances input to computations of probability. Thus, Griffin's success in the present experiments are also consistent with parrots' competence with modal reasoning as well as propositional, deductive reasoning.

There is, however, a leaner interpretation of Griffin's success. Importantly, there is no logical necessity that the updating process must be a deductive inference. Models of visual working memory, for example, are not likely in propositional format. Rather, such models are likely iconic representations, consisting of one symbol for each individual in a small set (see Carey, 2009, for review). Negation within such mental models could be realized by "option elimination procedures", and such procedures could well be domain-specific computations tailored to specific kinds of models and particular goals. For example, in the context of search, crossing out an option upon encountering evidence of an empty location need not require a mental symbol for *not*. Furthermore, it is known that infants, at least, can create models of two small sets, keeping these representations separate from each other, updating one without impacting the representation of the other and even combining them hierarchically to circumvent the limits on parallel individuation within working memory (Halberda et al., 2006; Rosenberg & Feigenson, 2013). It seems at least plausible that this mental model machinery is all that is needed for Griffin's success on these tasks. It would be necessary, however, to specify how he keeps the models of each side separate, and how he appropriately updates each side as he subsequently receives more information. With respect to this alternative explanation based on visual mental models, whatever updating computations are available to support probability attributions, they must yield the same pattern of success that propositional reasoning would yield in these tasks.

Similarly, the representations of possibility needed for success on the 3- and 4-cups tasks may not require explicit modal concepts such as "possible", "necessary", "impossible". Probability representations are needed for all learning, and can be realized by representing strengths of association, or frequencies (e.g., see Rescorla-Wagner models of associative learning) and thus do not require explicit representations of possibility as such. Indeed, recent experiments reveal striking failures by apes and young children at explicit representations of possibilities that can condition behaviour. For example, chimpanzees, orangutans, and 2 1/2-year-old human children fail to prepare for two equally likely possibilities when trying to catch a ball whose trajectory is uncertain. Redshaw & Suddendorf (2016) showed subjects a tube with a split at the end so that it looked like an upside-down Y (see Figure 1 in Redshaw & Suddendorf, 2016) and demonstrated that a ball dropped in the top came out randomly, and equally often, from either of the two legs of this upside-down Y. They then allowed the apes and children to try to catch the ball. Children under 3 years of age and the apes almost always put out one hand, guessing on each trial. Even over 12 trials, very few children and only one ape covered both possibilities with their two hands, and then not on all trials. It was not until age 4 that substantial numbers of children spontaneously put out both hands on the first trial. Children of age 2 1/2 never learned to do so, even after 12 trials of trying to catch the ball, and apes did so only after many extra trial blocks. Success that occurred after feedback could be accounted for by learning a rule "cover all the openings" rather than by evaluating possibilities. These failures suggest either that apes and humans less than 3 years old cannot entertain two possibilities, represented as such, at once, and cannot easily condition a response based on such a representation, or at least do not do so spontaneously in this task.

Additional studies are needed in order to decide between a mental model account of Griffin's success versus an account that draws on representations of propositions, including concepts of logical connectives and modality. One such study might be to search for evidence for negation functions that do not involve crossing out an option — for example, could Griffin learn two rules, that X predicts something on the left and that anything that is not an X predicts something on the right? Interestingly, several attempts to find such behaviour in human infants have failed — they learn the affirmative rule but not the negative one (e.g., Feiman et al., 2015; Hochmann et al., 2018). Similarly, could Griffin learn a rule conditioned on the relation of exclusive *or*? Could a study be designed to test whether Griffin can, like 4-year-olds but unlike 2 1/2-year-olds, prepare actions that simultaneously cover two mutually exclusive possibilities? Current work with Griffin is pursuing all of these lines of study.

8.2. What representational capacities underlie Griffin's out-performance of both apes and pre-school children?

Both the 3- and 4-cups tasks make substantial demands on working memory. The rewards were hidden in two distinct events, such that the participant must remember what was hidden on each side of the overall array in order to represent the possible locations of food (Griffin) or stickers (preschoolers) when planning a choice. As mentioned in the introduction, Griffin's working memory is markedly better than preschoolers' and likely that of chimpanzees (review in Reid, 2008); indeed, his updating working memory for mental models of what is hidden is equal to that of human adults (Pepperberg & Pailian, 2017). Thus, it is certainly likely that his better working memory provides at least a partial account for this better performance.

The hypothesis that Griffin's greater working memory capacity is the *complete* explanation for his better performance is consistent with the hypothesis that what underlies the developmental improvements between ages 2 1/2 and 5, and the comparative difference between apes and human adults, is *solely* improvements in working memory over evolution and ontogenesis. On this hypothesis, apes and very young preschoolers (2 1/2- and 3-year-olds) distinguish certainty from mere possibility, and establish certainty on the basis of the disjunctive syllogism. On this hypothesis, children's low level of correct choices on the 3- and 4-cups tasks and failure of the 2 1/2-year-olds on the 4-cups task, as well as apes' failures to display evidence for deductive reasoning in the Call & Carpenter (2001) task, solely reflects performance limitations due to the demands of updating working memory on these tasks.

There are, however, several reasons to doubt this hypothesis. We note that the working memory demands of the 4-cups task are markedly higher than those of the 3-cups version; in addition to there being one more cup, the task requires updating a working memory model upon being shown that one of the cups is empty. Consequently, if the main limit on correct choices derives from poor working memory, we would expect to see worse performance on the 4-cups task for all the children. Only for the 2 1/2-year-olds was this true. Even though the older children are far from ceiling on either task there is *no cost* of inference or of the greater working memory demands of the 4-cups version; performance is identical between the two tasks (on the 3-cups task, performance is between 60% choice of the certain cup at age 3 and 72% at age 5; on the 4-cups tasks performance is 58% at age 3 and 76% at age 5; Mody & Carey, 2016). This fact strongly implicates some other change, in addition to increasing working memory, in the developmental improvement that is observed on both tasks, at least between ages 3 and 5.

Establishing the basis for this change would be a source of hypotheses concerning the differences between great apes and children age 4 or older as well. Results from several tasks (e.g., Mody, 2016; Mody & Carey, 2016; Redshaw & Suddendorf, 2016) suggest that the poor performance of young preschool children on the 3- and 4-cups tasks is due not only to limited working memory, but also to problems in representing the related distinctions between knowledge and ignorance, or necessity and possibility, or probability of 1 versus probability of 0.5. And again, apes pattern with young children on these tasks (e.g., Call & Carpenter, 2001; Hanus & Call, 2014).

Young children and nonhuman primates surely have some implicit access to states of ignorance. As reported above, the 2 1/2-year-olds in Call & Carpenter's (2001) three tubes experiments looked into the tubes when they hadn't witnessed where the object was placed approx. 70% of the time, compared to approx. 40% of the time when they had seen the hiding; for apes the respective percentages are approx. 45% versus approx. 19%. Relatedly, young children even have explicit (verbalizable) access to states of ignorance. If shown a closed box they have not previously seen and asked if they know what is inside, even very young 3-year-olds answer “no”. However, further experiments show that preschool children’s explicit concepts of knowledge and ignorance are far from adult-like. If shown two toys (e.g., a frog and a lion), told that one of these toys will be put into the closed box while the box is out of their sight, and subsequently asked if they will know what is in the box, 100% of 3- to 4-year-olds answer, “yes, a frog” or “yes, a lion” (Rohwer et al., 2012). That is, young preschool children distinguish full ignorance from knowledge, but not partial ignorance from knowledge. At least in some circumstances when children make a mental model of a possibility, they do not understand that this model represents a mere possibility, distinguished from a certainty. The Hanus & Call (2014) experiment cited above testing the ability to distinguish certainty from mere possibility (choice of the container with one reward versus choice of one of two containers having one reward) revealed apes to be at the level of 2 1/2-year-olds. A high priority would therefore be to devise and carry out an identical task that can be given to young children, Griffin, and apes that taps representations of certainty versus mere possibility.

Rohwer et al.’s (2012) argument that children do not explicitly distinguish probability from certainty makes sense of two salient aspects of the youngest children’s and apes’ performance on the 3-cups task and its equivalent: why they are above chance (33%, assuming each cup a possibility) and also why they are so far from ceiling. Across every experiment to date on the 3-cups task (Mody & Carey, 2016; Andreuccioli, 2018; Halliday et al., in preparation), 2-year-olds and young 3-year-olds pick the certain cup only around 50–60% of the time. On the related task, apes score similarly (44%, Hanus & Call, 2014). If subjects guess which cup the reward is in within the two-cup side (just as they guess that it is a frog or a lion in the closed box), then they have a mental model that has one reward in the singleton cup and one reward in a particular cup from the other set. If they fail to represent the latter as mere possibility (essentially ignoring the other cup on the two-cup side), then they would choose one of the cups in their model randomly, leading to $\sim$50% choice of the singleton cup. This is exactly the data observed among very young preschoolers and apes.

Griffin’s data contrast markedly from those of young preschool children and apes. Rather than choosing the certain cup 50% of the time on the 3-cups task, he is near or at ceiling on this task in Experiments 1 and 5, and on the 4-cups tasks as well, somewhat outperforming even the 5-year-olds in

Mody & Carey (2016). Again, the apes' and young preschoolers' behaviour is consistent with the Rowher et al. analysis above, whereas Griffin's success, in contrast, shows him able to distinguish among probabilities when making choices in situations where preschool children and apes fail.

9. Concluding remarks

Although human adults most probably deploy the disjunctive syllogism when solving the 2-cups task, whether given the Premacks' or Call's version, success on the 2-cups task does not provide strong evidence for deductive reasoning, or for modal concepts like "necessity" and "possibility", or epistemic concepts like "certainty". That this is so is shown by young children's and apes' success at the 2-cups task in the face of their failures to display full knowledge of those concepts in other tasks, such as the 3- and 4-cups tasks (Hanus & Call, 2014; Mody & Carey, 2016), the knowledge and ignorance tasks (Call & Carpenter, 2001; Rowher et al., 2012), and the preparing-for-possibilities task (Redshaw & Suddendorf, 2012).

Griffin, in contrast, succeeds at the 3- and 4-cups tasks. What, if anything, does Griffin have, besides greater working memory, that 2 1/2-year-olds and chimpanzees lack, that is responsible for his success? These cups tasks most probably make demands on executive function beyond those of working memory. Impulsivity is often cited as a measure of executive function, and although the exact same tasks to test impulsivity have not been given to apes, Grey parrots, and preschool children, various studies compare these groups in pairs. For example, a study of behavioural inhibition in the context of delayed gratification (as in Mischel's [1974] marshmallow task) found that Griffin would wait 15 min for a favoured reward, over an immediate less favoured one (Koepke et al., 2015), whereas only a small percentage of the 4-year-olds in the standard task wait even 10 min to get 2 marshmallows rather than 1. In a different type of delayed gratification task, in which subjects had to travel further for a greater amount of reward, 6-year-old children outperformed chimpanzees and 3-year-old children were in between the apes and older children, although not statistically differently from either (Hermann et al., 2015). Grey parrots outperform apes in the ephemeral choice task, in which subjects are given two alternatives, A and B, with similar reinforcement provided for each; however, if they choose A, they also receive the reward associated with B, whereas if they choose B, the reward for A is removed. Thus, choosing A gives them two rewards, whereas choosing B gives them only one. Grey parrots succeeded on the task in fewer than 30 trials and in the reversal of the conditions in fewer than 50 trials (Pepperberg & Hartsfield, 2014); all but two chimpanzees failed to perform above chance in the initial 100 trials, taking 60 and 70 trials to learn, and then failed on the reversal (Salzwiczek et al., 2012). Only under significantly altered conditions did three orangutans learn the task, and only one of these learned the reversal in 80 trials. Some researchers (e.g., Zentall & Case, 2018) argue that optimal performance on this task depends on reducing impulsive choice. A full exploration of the executive function capacities of nonhumans and young humans would require identical paradigms that can be used across species, and would require expanding the set of executive functions studied from updating working memory and behavioural inhibition to other executive functions as well, such as response inhibition, set shifting, and sustained attention. Combined with studies of the role of executive function in any given task, such studies would contribute to an understanding of the cognitive factors that contribute to the species differences documented here. Such studies might reveal that Griffin's executive functions outstrip those of apes and pre-school children even more broadly than the above reviewed studies suggest. If they do, two questions arise. First, is this because he's a parrot or is this because of the over 20 years of schooling he has received? Executive functions are trainable (Bodrova & Leong, 2006). Comparative studies could be performed on adult parrots that lack 20 years of training during which they have learned that, in their environment, their behaviours can have predictable outcomes, and where figuring out seemingly arbitrary, non-ecological, rules guarantee those outcomes.

Whatever the basis for Griffin's success on the 3- and 4-cups tasks, it clearly does not require command of the language of logical connectives and modality; Griffin does not know that language. Whether we face the task of explaining how Griffin came to have the concepts of logical connectives and modality depends upon our further investigations that seek to adjudicate between a non-propositional mental model account of Griffin's success on these tasks and a logical account. If naïve parrots also succeed at these tasks, we would conclude that Griffin's over 20-year experience in the lab is not the key to parrots' success, and we would look to an ecological account for why parrots might have capacities apes and young children lack. However, not only might Griffin's training regime exercise and develop executive function, it also might specifically exercise and develop the logical capacities at issue in these studies. Griffin experiences very few trials tapping any probed construct in any one day (e.g., Pepperberg & Nakayama, 2016), and so he is never rewarded for adopting a strategy that yields "enough" rewards. Rather, he must provide a single answer, and there is a high premium on figuring out a strategy that is certain to succeed in any given context.

The present studies converge with others showing Griffin succeeds on non-linguistic tasks that draw on representational capacities not yet demonstrated by great apes and preschool children. These findings raise many questions; the search for their answers is just beginning.

Acknowledgements

We thank Caitlin Andrews, Lauren DiNicola and Leor Shuflita for helping to run many of the trials. This research was supported by a James S. McDonnell Foundation Grant to Susan Carey and donors to The Alex Foundation (particularly the Sterner Family, The Sunset Foundation, the Marc Haas Foundation, Janice Boyd, Joseph Golden, Alex and Michael Shuman, Nancy Chambers, Kathryn and Walter McAdams, Anita Keefe, Megumi Oka and the Makioka Foundation, Nancy Sondow, Katie Dolan, Deborah Rivel Goodale/Wildtones, The Robert D. Goodale Fund [via the Indian River Community Foundation], Elva and Bob Mathieson, Mauricio Catelli, Janine Kilty, Rose and Ken Lawlor, George and Lisa Bono/The Platinum Parrot, Arlene Levin Rowe/Grey Parrot Studios, Daisy Paradis/The Howard Bayne Fund, Dan Vollmer, Jana Julleis, Scott and Roxanne Bok, John Troth, the Dutch Parrot Foundation (Stichting Papegaai), Phoenix Landing, Michael Smith and LC Boros, The South Bay Bird Society, Don and Grace Wheeler, Pat Hill, Roni Duke, the Medwick Foundation, Bill Broach in memory of Nancy Clark, the Andrew De Mar Family Foundation, Ann Dey, Jan and Jeff Strong, James Ligman, Mark Hurley, the Central Indiana Cage-Bird Club, Gayle and Morton Speck, Susan Brown and James Armstrong). We thank Harrison's Bird Food for food and the Lafeber Company and Gini's Birdie Bread for treats. The study procedures comply with the current laws of the country under which they were performed (the United States).

References

Andreuccioli, L. (2018). Young children's implicit awareness of epistemic states: distinguishing certainty from uncertainty. — Unpublished BA Dissertation, Bath University, Bath.

Bloom, L., Lahey, M., Hood, L., Lifter, K. & Fiess, K. (1980). Complex sentences: acquisition of syntactic connectives and the semantic relations they encode. — J. Child Lang. 7: 235-261.

Bodrova, E. & Leong, D.J. (2006). Tools of the mind. — Pearson, London.

Call, J. (2004). Inferences about the location of food in the great apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, and *Pongo pygmaeus*). — J. Comp. Psychol. 118: 232-241.

Call, J. (2006). Inferences by exclusion in the great apes: the effect of age and species. — Anim. Cogn. 9: 393-403.

Call, J. & Carpenter, M. (2001). Do apes and children know what they have seen? — Anim. Cogn. 4: 207-220.

Carey, S. (2009). The origin of concepts. — Oxford University Press, New York, NY.

Clements, K.A., Gray, S.L., Gross, B. & Pepperberg, I.M. (2018). Initial evidence for probabilistic reasoning in a Grey parrot (*Psittacus erithacus*). — J. Comp. Psychol. 132: 166-177.

Davidson, D. (1982). Rational animals. — Dialectica 36: 317-327.

Descartes, R. (1637/1985). Discourse on the method. — In: Descartes: selected philosophical writings (Cottingham, J., Stoothoff, R. & Murdoch, D., trans.). Cambridge University Press, Cambridge.

Feiman, R., Carey, S. & Cushman, F. (2015). Infants' representations of others' goals: representing approach over avoidance. — Cognition 136: 204-214.

Feiman, R., Mody, S., Sanborn, S. & Carey, S. (2017). What do you mean, no? Children's comprehension of logical "no" and "not". — Lang. Learn. Devel. 13: 430-450.

Fodor, J. (1975). The language of thought. — Crowell, New York, NY.

French, L.A. & Nelson, K. (1985). Young children's knowledge of relational terms: some ifs, ors, and buts. — Springer, New York, NY.

Gathercole, S.E., Pickering, S.J., Ambridge, B. & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. — Dev. Psychol. 40: 177-190.

Giret, N., Miklósi, A. & Kreutzer, M. (2009). Use of experimenter-given cues by African Grey parrots (*Psittacus erithacus*). — Anim. Cogn. 12: 1-10.

Grether, W.F. & Maslow, A.H. (1937). An experimental study of insight in monkeys. — J. Comp. Psychol. 24: 127-134.

Griffin, D.R. (2001). Animal minds. — University of Chicago Press, Chicago, IL.

Halberda, J., Sires, S.F. & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. — Psychol. Sci. 17: 572-576.

Hanus, D. & Call, J. (2014). When maths trumps logic: probabilistic judgements in chimpanzees. — Biol. Lett. 10: 20140892. DOI:10.1098/rsbl.2014.0892.

Hedges, S.B., Parker, P.H., Sibley, C.G. & Kumar, S. (1996). Continental breakup and the ordinal diversification of birds and mammals. — Nature 381: 226-229.

Hermann, E., Misch, A., Hernandez-Lloreda, V. & Tomasello, M. (2015). Uniquely human self-control begins at school age. — Devel. Sci. 18: 979-993.

Hill, A., Collier-Baker, E. & Suddendorf, T. (2011). Inferential reasoning by exclusion in great apes, lesser apes, and spider monkeys. — J. Comp. Psychol. 125: 91-103.

Hill, A., Collier-Baker, E. & Suddendorf, T. (2012). Inferential reasoning by exclusion in children (*Homo sapiens*). — J. Comp. Psychol. 126: 243-254.

Hochmann, J.R., Carey, S. & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to learn a rule predicated on the relation different. — Cognition 177: 49-57.

Koepke, A., Gray, S.L. & Pepperberg, I.M. (2015). Delayed gratification: a Grey parrot (*Psittacus erithacus*) will wait for a better reward. — J. Comp. Psychol. 129: 339-346.

Lust, B. & Mervis, C.A. (1980). Development of coordination in the natural speech of young children. — J. Child Lang. 7: 279-304.

Mikolasch, S., Kotrshal, K. & Schloegl, C. (2011). African Grey parrots (*Psittacus erithacus*) use inference by exclusion to find hidden food. — Biol. Lett. 7: 875-877.

Mischel, W. (1974). Processes in delay of gratification. — In: Advances in experimental social psychology, Vol. 7 (Berkowitz, L., ed.). Academic Press, New York, NY, p. 249-292.

Mody, S. (2016). The developmental origins of logical inference: deduction and domain generality. — Unpublished PhD dissertation, Department of Psychology, Harvard University, Cambridge, MA.

Mody, S. & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. — Cognition 154: 40-48.

Morris, B.J. (2008). Logically speaking: evidence for item-based acquisition of the connectives AND & OR. — J. Cogn. Dev. 9: 67-88.

Penn, D.C., Holyoak, K.J. & Povinelli, D.J. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. — Behav. Brain Sci. 31: 109-130.

Pepperberg, I.M. & Carey, S. (2012). Grey parrot number acquisition: the inference of cardinal value from ordinal position on the numeral list. — Cognition 125: 219-232.

Pepperberg, I.M., Gray, S.L., Lesser, J.S. & Hartsfield, L.A. (2017). Piagetian liquid conservation in Grey parrots (*Psittacus erithacus*). — J. Comp. Psychol. 131: 370-383.

Pepperberg, I.M. & Hartsfield, L.A. (2014). Can Grey parrots (*Psittacus erithacus*) succeed on a "complex" foraging task failed by nonhuman primates (*Pan troglodytes*, *Pongo abelii*, *Sapajus apella*) but solved by wrasse fish (*Labroides dimidiatus*)? — J. Comp. Psychol. 128: 298-306.

Pepperberg, I.M., Koepke, A., Livingson, P., Girard, M. & Hartsfied, L.A. (2013). Reasoning by inference: further studies on exclusion in Grey parrots (*Psittacus erithacus*). — J. Comp. Psychol. 127: 272-281.

Pepperberg, I.M. & Nakayama, K. (2016). Robust representation of shape in a Grey parrot (*Psittacus erithacus*). — Cognition 153: 146-160.

Pepperberg, I.M. & Pailian, H. (May 24, 2017). Evolution of mechanisms underlying visual working memory manipulation: when "bird-brain" is a compliment. — Paper presented at the annual conference of the Vision Science Society, Tampa-St. Petersburg, FL.

Pepperberg, I.M. & Wilcox, S.E. (2000). Evidence for mutual exclusivity during label acquisition by Grey parrots (*Psittacus erithacus*)? — J. Comp. Psychol. 114: 219-231.

Pepperberg, I.M. & Wilkes, S. (2004). Lack of referential vocal learning from LCD video by Grey parrots (*Psittacus erithacus*). — Interact. Stud. 5: 75-97.

Pepperberg, I.M., Willner, M.R. & Gravitz, L.B. (1997). Development of Piagetian object permanence in a Grey parrot (*Psittacus erithacus*). — J. Comp. Psychol. 111: 63-75.

Péron, F., John, M., Sapowicz, S., Bovet, D. & Pepperberg, I.M. (2013). A study of sharing and reciprocity in Grey parrots (*Psittacus erithacus*). — Anim. Cogn. 16: 197-210.

Péron, F., Thornburg, L., Gross, B., Gray, S.L. & Pepperberg, I.M. (2014). Human–Grey parrot (*Psittacus erithacus*) reciprocity: a follow-up study. — Anim. Cogn. 17: 937-944.

Piaget, J. & Inhelder, B. (1975). The origin of the idea of chance in children (Leake, L., trans.). — W.W. Norton, New York, NY.

Premack, D. & Premack, A.J. (1994). Levels of causal understanding in chimpanzees and children. — Cognition 50: 347-362.

Redshaw, J. & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. — Curr. Biol. 26: 1758-1762.

Reid, D.W. (2008). Working memory: a cognitive limit to non-human primate recursive thinking prior to hominid evolution. — Evol. Psychol. 6: 676-714.

Rescorla, M. (2009). Chrysipus's dog as a case study in non-linguistic cognition. — In: The philosophy of animal minds (Lurz, R., ed.). Cambridge University Press, New York, NY, p. 52-71.

Rohwer, M.A., Kloo, D. & Perner, J. (2012). Escape from meta-ignorance: how children develop an understanding of their own lack of knowledge. — Child Dev. 83: 1869-1883.

Rosenberg, R.D. & Feigenson, L. (2013). Infants hierarchically organize memory representations. — Dev. Sci. 16: 610-621.

Salwiczek, L.H., Prétôt, L., Demarta, L., Proctor, D., Essler, J., Pinto, A.I., Wismer, S., Stoinski, T., Brosnan, S.F. & Bshary, R. (2012). Adult cleaner wrasse outperform capuchin monkeys, chimpanzees, and orangutans in a complex foraging task derived from cleaner-client reef fish cooperation. — PLoS ONE 7: e49068.

Schmitt, V. & Fischer, J. (2009). Inferential reasoning and modality dependent discrimination learning in olive baboons (*Papio hamadryas anubis*). — J. Comp. Psychol. 123: 316-325.

ten Cate, C. & Healy, S.D. (eds) (2017). Avian cognition. — Cambridge University Press, Cambridge.

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

Tornick, J.K. & Gibson, B.M. (2013). Tests of inferential reasoning by exclusion in Clark's nutcrackers (*Nucifraga columbiana*). — Anim. Cogn. 16: 583-597.

Voelter, C.J. & Call, J. (2017). Causal and inferential reasoning in animals. — In: APA handbook of comparative psychology (Call, J., Burghardt, G., Pepperberg, I.M. & Zentall, T.R., eds). American Psychological Association Press, Washington, DC, p. 643-671.

Watson, J.S., Gergely, G., Csanyi, V., Topal, J., Gasci, M. & Sarkozi, Z. (2001). Distinguishing logic from association in the solution of an invisible displacement task by children (*Homo sapiens*) and dogs (*Canis familiaris*): using negation of disjunction. — J. Comp. Psychol. 115: 219-226.

Zentall, T.R. & Case, J. (2018). The ephemeral-rewards task: optimal performance depends on reducing impulsive choice. — Curr. Dir. Psychol. Sci. 27: 103-109.

Appendix: Procedure formalization

We began the study expecting to be able to use procedures similar to the previous exclusion study (Pepperberg et al., 2013); that is, by simply increasing the number of cups on the tray to replicate Mody & Carey (2016). We thus began in 2014 by placing Griffin on a T-stand and presenting a green felt-covered tray familiar from the previous exclusion study, but now containing four differently coloured cups (blue, yellow, green, orange). Yellow and green cups had been used in the previous exclusion study; green and orange cups had been used in two intervening studies on prosociality (Péron et al., 2013, 2014), where green had a positive connotation and orange a negative one. Probably as a consequence of these prior studies, Griffin now showed a strong bias for choosing the green cup, and increasing the spacing between cups did not improve his behaviour. We next added a barrier between the cups to demonstrate that the four cups were actually two sets. Griffin then demonstrated a strong preference for one side, which we later learned (see below) was likely caused by an injury to one foot, affecting his balance and ability to turn and bend in one direction versus the other. If given a correction trial after erring, however, he did make an effort to choose correctly. During trials, when not favouring a particular position, he also seemed biased toward the blue cup, a colour never used in any previous trials. We tried switching to identical silver cups with white cardboard covers. All possible choices were now identical. Griffin's performance seemed to improve (i.e., he appeared to exhibit no bias to a specific cup), but he was reluctant to reach toward the outer two cups and a hiatus occurred when a research assistant graduated and needed to be replaced.

During this delay, Griffin seemed to have recovered his balance, and we then performed Experiment 1. We used knowledge gained from the previous protocols to reconfigure the task so as to avoid overlap with other studies involving use of both coloured plastic cups and the familiar green tray. We designed a protocol for the 4-cups task using materials described in 2. General methods. The white tray, to which Griffin was habituated from a task on liquid conservation (Pepperberg et al., 2017), had never been associated with either the silver or the coloured plastic cups and would thus allow us to differentiate the present protocol from all the previous trials. Use of the covered silver cups was retained from the earlier trials, but we now added the coloured pompons to individualize each cup as in Mody & Carey (2016).

Our plan was now (as of 2015) to return to the four-cup task; however, we found a huge bias for the cup in the third position; after finding the same preference on several three-cup retrials, we realized that he had now seriously (re-)injured his foot and that his balance was badly compromised. We tried placing him on a flat surface so he could walk to his choice, but he refused to wait in place until we performed the manipulations. We erected a barrier of cardboard and clear plastic during the manipulation, but that seemed to distort his vision. We therefore put the study on hold until his foot was completely healed, and proceeded with other experiments in the interim that did not require physical choice. Once Griffin was completely healthy, during the summer and fall of 2016 and into the winter of 2017, we performed Experiments 2 through 5.