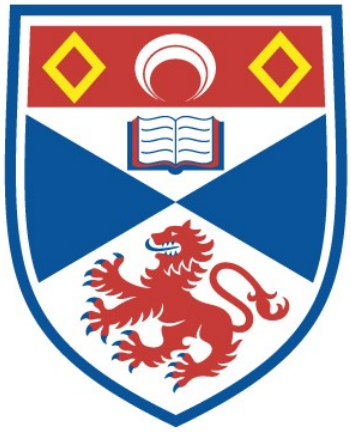
A comparative perspective on reasoning about possibility

Benjamin Jones

A thesis submitted for the degree of PhD
at the
University of St Andrews



2025

Candidate's declaration

I, Benjamin Jones, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 47,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2021.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 10/09/2024 Signature of candidate:

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 10/09/2024 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Benjamin Jones, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 10/09/2024 Signature of candidate

Date 10/09/2024 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Benjamin Jones, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date 10/09/2024 Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date 10/09/2024 Signature of candidate

Date 10/09/2024 Signature of supervisor

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Josep Call, for his invaluable guidance, patience and feedback throughout my PhD. I am truly grateful for him placing his confidence in me, allowing me the freedom to explore and develop my ideas while providing the structure necessary to see them through. I would also like to thank my second supervisor, Professor Amanda Seed, for mentorship and guidance through the process and insightful feedback on whatever problem I brought to her.

I am deeply grateful to the GRAPES lab for their collaboration, encouragement, and insightful discussions. Their diverse specialisms and constructive feedback greatly enriched my research projects and development as a scientist. In particular to Sarah Koopman, whose invaluable perspectives, wealth of experience, and thoughtful advice in navigating academia were always appreciated, and regular updates on royal gossip helped pass long, chimp-less research sessions.

Outside of our lab, I would like to thank Professor Sarah Beck, for several long discussions about counterfactual reasoning and for enthusiastic support in developing (and participating in) a study of regret in great apes. Importantly, to Nicole Furgala, who “doesn’t really care about logic”, but is fantastic at linking it to other areas of primate cognition and gently reminds me to make my writing sound less robotic.

Regarding research, I wish to thank Kate Grounds, Donald Gow, Callum Gibson, and the Edinburgh keeper team, whose tireless care for the chimpanzees and practical support make research at the BRU both possible and enjoyable. At Twycross, I am grateful to Lisa Gillespie and the research team for facilitating the setup of a new project. I want to thank Amanda Addison and the great ape team for not only enabling my research through practical support, advice and care for the apes but also for making me feel welcome during the many months I spent living at the zoo. For their assistance with research and coding, I would like to thank Eloise Dallas and Sadie Tenpas; and Eliza Colley-Illueca for her many hours of reliability coding and her humorous but inaccurate judgements of the chimpanzees’ personalities. Finally, to the apes, attentive and less so, who swapped their time for grapes and made this research possible.

I would like to express my thanks to my family and friends, some already mentioned, who have provided me with support, encouragement and a welcome distraction from academia. However, my final and dearest thanks go to my partner Sasha, who’s unwavering and unconditional support has made finishing this thesis possible.

Funding

This work was supported by the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013)/ERC grant agreement no. 609819, SOMICS, Templeton World Charity Foundation, grant ID: TWCF0314.

This work was supported by St Andrews School of Psychology and Neuroscience.

Research Data Access Statement

Research data underpinning this thesis are available at https://doi.org/10.17630/8cda57d5-64ed-43a1-b9c6-f5c7e2c94fa8

Thesis Abstract

As humans we spend considerable effort contemplating possibilities, whether planning for the future or pondering on the past, reasoning about what is possible is an essential aspect of our lives. For that reason, it is relevant to ask whether we are unique in our ability to consider multiple possible futures or pasts. Across four experimental chapters, I attempt to test some of the explicit claims laid out by the varying models of non-linguistic reasoning. In Chapter 2, I aimed to test Leahy and Carey's (2020) minimal model of possibility. I developed a novel paradigm, post-decision wagering, and used it to demonstrate that great apes reason about the existence of multiple, incompatible possibilities. In Chapter 3, I modified the post-decision wagering paradigm, to test whether great apes were able to reason via the disjunctive syllogism. Finding that, if given information about the unchosen cup, subjects were able to adaptively choose between their original choice and a fractional reward. In Chapter 4, I tested all four great ape species using Mody and Carey's (2016) classic 4-cup disjunctive syllogism task and Ferrigno et al.'s (2021) modification. Apes switched adaptively in-line with logical reasoning, but performance was poor compared to the 2-cup variant and, when I included additional control trials, subjects failed to choose above chance levels. Chapter 5 explored whether chimpanzees were curious about counterfactuals using a modification of Call and Carpenter's (2001) 3-tube paradigm. Showing that, after being given a choice between 2 of 3 differentially baited tubes, subjects were more likely to check the contents of the unchosen than the unavailable tube. Finally, Chapter 6 discusses how these findings contribute to our understanding of how great apes reason about possibility. I explore whether our data support any of the previously proposed hypotheses and why performance breaks down in the 4-cup tasks.

Table of contents

Index of tables

Index of figures

1. General Introduction

1.1 Thinking about uncertainty

Games of luck emerged thousands of years ago in ancient Egypt, yet even the Greeks, despite all of their advances in formal mathematics, believed that the cast of the die was in the hands of the gods (Bernstein, 1996). However essential it appears to be for modern life, even the most basic probability theory has existed for a remarkably short fraction of human history.

Knight (1921) proposed that the term risk should be used to describe situations when the probabilities of each possible outcome were known and could be calculated, while uncertainty would be used to describe decisions where the subject was lacking information relevant to making the decision. Concurrently, Keynes (1921) reached a similar conclusion but describes *calculable* and *incalculable* risk, the latter being reserved for cases which there was no scientific way to rationally predict. Chua Chow and Sarin (2002) kept this knowledge based definition but added an additional level, *unknowable uncertainty*. Which they described as determined but not knowable, citing the example of guessing the number of seeds in an apple, neither the asker nor the answerer know the correct answer, yet no external source of knowledge can resolve the uncertainty.

Ellsberg (1961) instead describes risk and *ambiguity*. He shows that people prefer a choice under risk, an urn containing 50 red and 50 black balls, over a choice under ambiguity, an urn with an unknown quantity of red and black balls[1]. Kahneman and Tversky (1982) make a comparable division between internal uncertainty, which is caused by a lack of knowledge, and external uncertainty which is caused by randomness[2]. Crucially for Kahneman and Tversky, as it was for Knight (1921), external uncertainty can be modelled, while internal uncertainty, cannot. Under this distinction, external uncertainty is caused by systems, which have

---

[1] This paper is a challenge to subjective expected utility theory (Savage, 1954; cited in Camerer and Weber, 1991), which denotes that people assign subjective probabilities to their beliefs about states of the world and choose rationally based on these probabilities. When the subjective probability of each distribution of balls in urn 2 (1-99 red balls) is combined with the likelihood of drawing a red ball under each of these distributions the result is p(red) = 0.5. Therefore, the subject should be indifferent, yet they are not.

[2] Howell and Burnett (1978) also refer to internal and external uncertainty, but instead specify these terms as relating to the locus of control, the agent can control their internal level of uncertainty by searching for more information, but not their external uncertainty. Reminiscent of reducible and irreducible uncertainty, which are used in computer science.

dispositions to produce outcomes, these outcomes can be projected distributionally based on past outcomes for repeated occurrences, or by one off estimation based on the propensities of the components of the system. Kahneman and Tversky also make a temporal distinction, proposing that "*Uncertainty about past events is likely to be experienced as ignorance, especially if the truth is known to someone else, whereas uncertainty about the future is more naturally attributed to the dispositions of the relevant system*" (p152, 1982).

The final noteworthy distinction is that of fundamental uncertainty (Dequech, 2000), which covers the occurrence of an event that you cannot even imagine, so is the only *true* uncertainty. While all else should be considered ambiguity – uncertainty created by missing probabilities which is reducible prior to the point of decision. In the Ellsberg paradox, we are told that there are only red and black balls within the second urn, therefore it is possible to populate the probability distribution even if we are not able to ascribe the possibilities. Clearly, this final definition does not produce any hypotheses that are testable under a comparative or developmental setting.

Table 1.1.1: Relating the different historical concepts of uncertainty.

	Risk [1]	*Ambiguity* [2]	*Uncertainty* [1]	*Fundamental Uncertainty* [3]
Outcome space	Known	Known	Known	Impossible to imagine
Probability Distribution	Known	Unknown	Unknown	Impossible to assign
Reducibility	Irreducible, but modellable	Reducible by information seeking	There subject has a fundamental lack of data about the system itself.	The necessary information does not exist.
Locus of the uncertainty [4]	External – Caused by dispositions of the system	Internal – Caused by ignorance	Internal – Caused by lack of information.	The necessary information does not exist.
Example	Rolling a fair dice	Ellsberg's 2nd urn.	Will war break out in Europe.[1]	Radical technological innovation.[3]

[1](Knight, 1921), [2](Ellsberg, 1961), [3](Dequech, 2000). [4](Kahneman & Tversky, 1982)

In the modern literature, uncertainty is generally divided into epistemic uncertainty, which refers to uncertainty pertaining to a lack of knowledge, and aleatory (or physical) uncertainty, which refers to the uncertainty inherent to physical systems, the exact value of which exists

within a range, so can be modelled to an extent, but varies by chance (Thunnissen, 2003). A useful perspective comes from computer science and machine learning literature, which interchanges epistemic and aleatory uncertainty with reducible and irreducible uncertainty, respectively (Hüllermeier & Waegeman, 2021).

Studies in comparative and developmental psychology use the terms *epistemic* and *physical* uncertainty and simplify their meanings to *resolved but unknown uncertainty* and *unresolved uncertainty*, respectively. Throughout this thesis I will use a simplified definition of uncertainty, taken from Bedford and Cooke (2001): "*that which disappears when we are certain*". I will consider epistemic uncertainty to be outcomes which have been decided but their value is unknown to the subject, and physical uncertainty to be outcomes which are as yet undecided, as these are the definitions adopted by the comparative and developmental literature which I will be referencing.

Children under the age of 6 are more likely to make a guess as to the actual outcome when presented with epistemic uncertainty, prepare for multiple eventualities when facing physical uncertainty (Robinson et al., 2006). Unlike adults, 6-year-old children prefer to guess the outcome of an already determined dice roll, as do 15-year-olds (Robinson et al., 2009). Adults, while adept at evaluating possibilities, are more likely to place a bet on a hypothetical outcome that has yet to be decided than one which has happened but they are unaware of (Brun & Teigen, 1990; Heath & Tversky, 1991; Robinson et al., 2009). Study participants also find a bet less attractive if it is also being evaluated by an individual with greater knowledge (Fox & Tversky, 1995), which the authors call the *comparative ignorance hypothesis*[3]. This is supported by the finding that when considering unknowable uncertainties, specifically the number of pips in an uncut apple, adults bet at levels more comparable to physical uncertainty (Chua Chow & Sarin, 2002). However, if given a semblance of agency over the chance outcome, through rolling the dice themselves, this preference for the undetermined outcome disappears (Harris et al., 2011; Robinson et al., 2009). Adults also speak about epistemic and physical uncertainty differently (Ülkümen et al., 2016), describing epistemic uncertainty using egocentric confidence statements but physical uncertainty using likelihood statements. Most recently, Fox, Goedde-Menke and Tannenbaum (2021) showed that subjects prefer a purely physical gamble over a mixed epistemic and physical gamble and prefer both over a purely epistemic gamble; their

[3] This is a development Heath and Tversky's (1991) *competence* theory, if you are gamble while incompetent then your success will be ascribed to luck, while if you gamble while competent then your success will be ascribed to skill. Epistemic uncertainty makes you consider what you could know but don't, eliciting feelings of incompetence and thus, there is no reputational upside to gambling but all of the downside.

conclusion being that the presence of an *aleatory hedge*, the possibility that you may be saved from a being incorrect epistemically by a fortuitous aleatory outcome, makes this gamble more attractive.

1.2 Reducing uncertainty.

Much as machine learning algorithms distinguish between reducible and irreducible uncertainty, biological organisms must attempt to predict their environment. The goal of all organisms is to reduce uncertainty so that they can attempt to predict the future which they will experience. Then, once they have experienced that future, they can revise their model of the world to better predict what they will experience next. This iterative process of making predictions and revising them in the face of new evidence is known as Bayesian inference. While the debate continues whether humans are Bayesians, it is the basis of much of the neuroscience literature on decision making, so is valuable to define.

In his theory of sentient behaviour Friston describes a process he terms *active inference* (Friston et al., 2009; Parr et al., 2022). As biological agents we move around the world constantly sensing our environment, but the brain does not have access to the actual world therefore it must use *inference* to make its best estimate as to what the actual state of the world is based on only the neuronal input it receives. This is a process akin to Bayesian inference, whereby the brain makes a prediction about the world, which it then updates with new evidence. How accurate its model of the world is can be measured by *surprise,* the difference between the predicted state of the world and the sensed state of the world, referred to in the model as *free energy*. Friston argues that when a certain element of the model does not make sufficiently concrete predictions the organism will search for more information, for example, through movement.

However, while this simple form of inference can describe how a simple organism makes sense of their environment, we are undoubtedly capable of much more advanced reasoning capabilities. Human reasoning can be split into 3 forms. Firstly, is *abductive* reasoning, the "process of forming an explanatory hypothesis" (Pierce, 1903, in Mcauliffe, 2015), this later became referred to as 'inference by best explanation' (Lipton, 2000) and while there is debate as to the nuanced differences between the two (Campos, 2011; Mcauliffe, 2015), both maintain that abduction simply makes a hypothesis which best explains the data and does not make predictions as to the consequences of that hypothesis. However, much more common in human reasoning are the two more advanced levels of reasoning, inductive and deductive. Like abductive reasoning, *inductive* reasoning relies on collecting data from the environment to make

a hypothesis, however, it additionally makes an expectation about the consequences of that hypothesis (Lipton, 2017). Finally, *deductive* reasoning uses strict rules or axioms, statements which are known to be universally true, to make new prepositions from initial premises (Douven, 2021). These logical arguments take the form of a syllogism, a set of premises which lead to a conclusion. Rules can be applied consecutively and repeatedly, but provided that the initial premises are valid, and that the logical rules are valid and applied correctly, then the new propositions can be considered truths. This final aspect is crucial, as the reasoner (or in an experimental setting the experimenter) is responsible for setting the premises, incorrect premises can lead to a conclusion which is deductively correct but causally inconsistent.

The *unitary* view (Rips, 2001) suggest that humans view inductive and deductive arguments on a spectrum. As deduction creates a dichotomy between deductively correct and deductively incorrect, a correct answer can only exist at the top end of a spectrum, however that same top end of a spectrum can also be held by an inductively strong argument. The prime example of which being gravity, because gravity is theoretically falsifiable, it remains an inductive argument, but one would place a high degree of certainty to a dropped marble being found at the bottom of a ramp, rather than suspended part way up. However, under the two-process argument, which the data provide more support for, the two are treated differently, with induction being measured in terms of similarity between the premise and the conclusion, while deduction being measured based upon validity (Heit & Rotello, 2010; Rips, 2001). Similar to the dual-process model in other areas of psychology, it is thought that fast judgments are made based on similarity, inductive reasoning, while slow judgements are made analytically based on deductive validity (Heit & Rotello, 2010). Evidence against an inductive inference is weighed more heavily than evidence for it and absence of evidence is treated as evidence of absence (Johnson et al., 2015). Schurz (2021) argues that induction must be the evolutionary ancient condition, because even a majority of adults do not intuitively understand some tenets of basic logic. This raises the question of whether the same advanced inference abilities are present in non-human primates, herein primates, without the affordances of deduction.

1.3 Reasoning in primates

In discussions of reasoning in animals, authors often turn to the 2nd century BC anecdote of Chrysippus' dog (De Waal, 2019; Engelmann, Haux, et al., 2023; Floridi, 1997; Rescorla, 2009), in which, during the pursuit of an animal which has fled down one of 3 paths, the dog sniffs the first, then the second, then goes down the third without sniffing (in Rescorla, 2009). Chrysippus suggests that once the dog has eliminated the first two paths as possibilities, the

third is the only option so stopping to sniff it is an inefficient waste of effort. To probe this anecdote, Watson and colleagues (2001) tested children aged 4-6 and domestic dogs (*Canis familiaris*) in their ability to use inference by exclusion in a stage 6 object permanence task. Defining logical search as follows:

> *"Under the guidance of only a logically based commitment to search (i.e., a belief that the object must be in one of the hiding places), each failure to find the object in a selected place amounts to an increase in the implied likelihood of the object being at a place not yet searched."(Watson et al., 2001, p. 221)*

Using a displacement device, the experimenters covertly deposited a target item behind one of three screens before showing the subject it was now empty. Their dependent measure was whether, after having failed to find the target behind the first 2 screens, whether subjects would accelerate when approaching the 3rd and final screen. The authors found that while the children did increase the speed with which they approached the final screen, the dogs did not. While generalising approach velocity as a universal measure of certainty has its limitations, the authors propose this as evidence of dogs and children younger than the age tested are not capable of inference by exclusion.

In a conceptually similar paradigm, Call and Carpenter (2001) tested great apes (3 orangutans (*Pongo pygmaeus)* and 12 chimpanzees (*Pan troglodytes*)), and 2-year-old children in a meta-cognitive paradigm that involved either 2 or 3 horizontal tubes placed at table height, one of which was baited with a target item. The subjects could move their head down to check the contents of the tubes before making their choice. While the study produced a number of findings, with regard to inference ability apes chose the last tube without checking it on 28% of 2-tube trials, and 13.9% of 3-tube trials. However, of the 11 chimpanzees who engaged with the 3-tube task, only 3 checked the tubes efficiently and from these 3 only 1 reliably stopped searching after finding 2 empty tubes, so her results make up a large proportion of the total. Nevertheless, because checking the final tube represents a low cost, checking it cannot truly be considered a failure of inference, particularly as only 2 of the chimpanzees reliably stopped searching once they had seen the food piece. The 2-year-olds were only tested on the 3-tube paradigm and generally were more efficient in their searches but inferred the contents of the last tube without checking it on only 4.6% of empty-empty-baited trials. Suggesting that they too lacked the ability for inference by exclusion. Similarly, tufted capuchins (*Cebus apellus*) searched until they found visual evidence, including when the tubes were transparent (Paukner et al., 2006).

Thus, the evidence presented so far would suggest that dogs, apes, monkeys, and children below three years of age do not conform to the behaviour of Chrysippus' dog, while children above the age of 4 do.

Alternatively, inference by exclusion can be directly tested using an object search task, as choice behaviour in line with rational thought can be considered evidence of logical reasoning. To illustrate this, Call's (2004) 2-cup 1-item task[4], herein 2-cup task, presented a subject, in his case great apes, with two upturned cups one of which contains a target item. When participants were given indirect evidence about the location of the item, either by the experimenter shaking one cup or showing them the empty cup, they were able to infer the location of the target item. Crucially, subjects were above chance from the first trial and were unable to learn to use arbitrary cues, either an audio recording or a tapping sound, which rules out associative learning as an explanation. The visual 2-cup paradigm (or a species suitable equivalent) has been used to illustrate inference by exclusion in primates (Bräuer et al., 2006; Call, 2006; De Petrillo & Rosati, 2020; Heimbauer et al., 2019; Hill et al., 2011; Marsh et al., 2015; Paukner et al., 2009; Petit et al., 2015; Schmitt & Fischer, 2009), birds (Danel et al., 2021; Mikolasch et al., 2012; O'Hara et al., 2015, 2016; Pepperberg et al., 2013), domestic ruminants (Duffrene et al., 2022; Nawroth et al., 2014[5]), domestic dogs, (Erdőhegyi et al., 2007), and children as young as 23-months (Mody & Carey, 2016). Crucially, the comparative literature consistently shows individual differences in inference ability, and some individuals may solve these tasks by other strategies (Schmitt & Fischer, 2009), for this reason some authors to instead use the term *exclusion performance* (e.g. Nawroth et al., 2014; Schloegl et al., 2009).

The auditory version of the 2-cup task (Call 2004) requires that subjects use *diagnostic inference*, inferring the contents of the cup via secondary clues. This can either be through confirming the consequent, realising that the cause of the sound when the cup is shaken is the presence of the object inside the cup, or denying the consequent, that the cause of the silence is the *absence* of an item inside the cup. Auditory diagnostic inference is distributed more unevenly[6], being present in great apes (Call, 2004; Hill et al., 2011), tufted capuchins, *Cebus*

---

[4] Adapted from Premack & Premack 1994

[5] These two references provide conflicting data regarding the inference abilities of domestic sheep. Nawroth et al. (2014) find that goats are capable of inference by exclusion while sheep are not, which they ascribe to the differences in their dietary flexibility. In contrast, Duffrene et al. (2022) found that sheep passed their inference by exclusion task, and attribute the difference to the sheep that they tested having been housed in a petting zoo, who's environment was more cognitively enriching than those tested by Nawroth et al. (2014), who were housed in a standard breeding facility.

[6] For the purpose of classification I am considering auditory diagnostic inference to be present if subjects pass either a full information condition (the experimenter shakes both containers) or if they pass both the

apella, (Sabbatini & Visalberghi, 2008), brown lemurs, *Eulemur fulvus* (Maille & Roeder, 2012), ruffed lemurs, *Varecia* spp, (De Petrillo & Rosati, 2020), grey parrots, *Psittacus erithacus*, (Schloegl et al., 2012), skuas, *Stercorarius antarcticus* (Danel, Rebout, Bonadonna, et al., 2022), sloth bears, *Melursus ursinus*, (Amici et al., 2017)and domestic pigs, *Sus scrofa domestica* (Nawroth & von Borell, 2015); but absent in black lemurs, *Eulemur macaco*, (Maille & Roeder, 2012) sifakas, *Propithecus coquereli*, (De Petrillo & Rosati, 2020), squirrel monkeys. *Saimiri sciureus*, (Marsh et al., 2015), olive baboons, *Papio anubis*, (Petit et al., 2015; Schmitt & Fischer, 2009), and in both dogs and wolves, *Canis spp*, (Lampe et al., 2017; Rivas-Blanco et al., 2024). The uneven distribution between closely related species would suggest that this may be a more advanced capacity, however, the sample sizes in many of these studies are particularly small so may not be representative and risk both type I and type II errors.

Notably, donkeys, *Equus asinus*, showed the reverse pattern, passing the auditory version but failing the visual (Danel et al., 2022; also see Maille and Roeder, 2012) and elephants, *Elephas maximus*, failed visual and auditory versions of the task but passed an olfactory equivalent (Plotnik et al., 2013, 2014). Both of which are their dominant modalities. Similarly, sloth bears, who feed largely on hidden insects, are able to pass both versions of the auditory diagnostic inference tasks (Amici et al., 2017). These examples are in-line with the suggestion those primate species' who excel in the auditory version of the task are more likely to be reliant on extractive foraging (Schmitt & Fischer, 2009). This is the core of the *adaptive specialisation hypothesis* (Krebs, 1990), whereby inter-species differences in cognition are driven by differences in dietary complexity.

Feasibly, this association between concealed food, shaking and sound, could have been a learned one, as zoo-housed subjects could learn this association by interacting with either full or depleted puzzle feeders provided as part of the cognitive enrichment provided to captive apes[7]. To support this suggestion, capuchins show improved performance in auditory test trials after being given training trials where they have the opportunity to manipulate baited containers (Sabbatini & Visalberghi, 2008). However, apes can also draw diagnostic inferences from novel stimuli, such as using a trail left behind when moving a cup to locate a hidden yoghurt pot

---

shake empty and *shake baited* partial information conditions, if subjects only pass the *shake baited* condition then they could pass by simply choosing the container the experimenter last manipulated (e.g. Lampe et al., 2017).

[7] This explanation could also be used for inferring baited contents from weight when lifting (Schrauf & Call, 2009, Hanus and Call 2011).

(Völter & Call, 2014a), or inferring that baiting was the cause of tipping of a novel see-saw apparatus (Hanus and Call, 2009).

Nevertheless, to solve Call's 2-cup task, an agent does not necessarily need use any logical process. Instead, they can merely mark the two locations as '*maybe A, maybe B*', when they are shown disconfirming evidence about A, they do not need to change their evaluation of likelihood B is correct, instead their revised model can be '*not A, maybe B*'. Crucially, this understanding does not require any comprehension of the logical operator OR, the two can be held as entirely separate models with no relation to one another. In an even simpler explanation of these findings, the subject need not have any expectation at all about the location of the food but in a forced choice paradigm, avoiding a known empty cup leads to choosing B, as it is the only remaining option. This has become known as the *avoid the empty cup* hypothesis (Paukner et al., 2006).

Call (2022) directly tested the *avoid the empty cup* hypothesis using a 3-cup paradigm by baiting a pair of cups, behind a barrier, while a third cup rested on the table untouched. He then showed the subject that one of the possibly baited cups was empty and gave the subject a choice between the 3. Nineteen of 23 great apes selected the cup that had been behind the barrier above chance levels and, as a group, subjects were above chance from the first trial. In a control condition where the experimenter did not reveal the contents of the empty cup, subjects were at chance. In a second experiment in which the experimenter instead removed the baited cup, the apes showed no preference for the cup behind the barrier during baiting, suggesting that subjects treated it equally to a cup which had zero chance of having been baited. These two experiments suggests that apes are solving this problem using inferential updating.

Notably, Premack and Premack's (1994) original paradigm on which the 2-cup task was based did not show a container as empty, instead the subject saw an experimenter baiting the two containers with different food items and then the subject saw the experimenter eating one of the two items, then had to infer that they had taken it from the baited container, so to select the other. While two younger individuals failed the task, one 10-year-old chimpanzee chose adaptively based on which fruit the experimenter was eating. Call (2006) later replicated this finding with a larger sample and a test for associative learning but found the same results. Likewise, other paradigms designed to account for avoidance of the empty cup have found positive results in corvids (Jelbert et al., 2015; Schloegl et al., 2009) and ruminants (Duffrene et al., 2022). Nevertheless, while evidence suggests that apes are capable of reasoning without this non-cognitive strategy, we must still consider it as a viable explanation for those species that has not been actively controlled for.

Mody and Carey (2016) sought to disambiguate a logical from an associative understanding of inference by exclusion, by first presenting 2.5-year-olds with a large format version of the 2-cup task, then 2.5- to 5-year olds with a novel 3-cup 2-item (3-cup) and a 4-cup 2-item (4-cup) search task, with the 3-cup task serving as training trials for the 4-cup task. In the 3-cup task the experimenter presented the child with 3 empty cups and a small screen which could cover 2 cups at a time. They then used the screen to covertly bait the pair with one sticker, and the single cup with another, counterbalanced by position and order. Then offered the subject a choice of which cup they would like to search. Notably, the 3 training trials were interspersed with 2 demonstrator trials, in which the second experimenter correctly chose the certain cup and explained why they were doing so. All groups chose above the 33% chance level, (2.5-year-olds: 47%, 3-year-olds: 60%, 4-year-olds: 71%, 5-year-olds: 72%.)

In the 4-cup task the authors presented children with 2 pairs of cups, each pair baited with 1 sticker behind the screen, as in the 3-cup task. This meant that each cup had a 50% chance of containing a sticker. The experiment was setup as a game where the subject would take turns with an experimenter to attempt to find a sticker. The experimenter went first and always failed to find a sticker, creating a situation analogous to the negative visual evidence condition (show-empty) in the 2-cup task, but only for one pair. The correct response is to choose the pair of the empty cup, as its likelihood of containing a sticker has gone from 50% to 100%, hence it is referred to as the target cup. If the subject simply avoided the empty cup, they would choose equally between the remaining 3, however, if they had inferentially updated their beliefs about the contents of the target cup, they would then choose it over the others. Two-and-a-half-year-olds, who had passed the 2-cup task, only chose the target cup on 36% of trials, all other groups chose above the 33% chance level (3-year-olds: 58%, 4-year-olds: 64%, 5-year-olds: 76%).

Thus, the authors argue that before the age of 3 children are unable to combine the logical operators *not* and *or* to reliably solve the disjunctive syllogism A *or* B *not* A, *therefor* B. The authors argue that the older groups who passed the task are in fact using logic rather than probabilistic updating because they "*chose the target cup just as often in test trials as they did in training trials, in which they could directly observe that a sticker was being hidden there*" (Mody and Carey, 2016, p. 46). While technically the children did not directly observe that the sticker was hidden there as it was also baited behind a screen, 3- to 5-year-old children chose at the same rate when they had reached a conclusion by logical inference as when they had reached it through object permanence.

In a comparable study, Hanus and Call (2014) tested 8 chimpanzees in an array of combinations of certain and uncertain choices, one of which was equivalent to the 3-cup task. Subjects chose

between a cup baited with one food item and a pair of cups one of which was baited with an identical food item. Like the youngest children, the chimpanzees chose the certain cup on 44% of trials and at the individual level, 7 of the 8 chimpanzees chose the certain cup on 50% of trials. The comparable findings between chimpanzees and young children, along with a second study, Redshaw and Suddendorf's (2016) forked tube paradigm, go on to form the basis of Leahy and Carey's (2020) minimal model of possibility.

Redshaw and Suddendorf's (2016) forked tube task[8] is a minimal way to test young children and great apes' ability to prepare for mutually exclusive possibilities in the absence of language. The authors fashioned a large vertical Y-shaped maze from drainpipe, inverted to have one entrance at the top and two-exits at the bottom. The task then simply involved dropping a suitably desirable item, a bouncy ball for the children or a grape for the apes, into the top of the apparatus and observing whether subjects covered both exits. The great apes, 3 chimpanzees and 5 orangutans, and 2-year-old children tended to cover only one exit, while most 4-year-olds and some 3-year-olds spontaneously covered both exits from the first trial. Crucially, after covering both openings on one trial, none of the 4-year-olds later regressed to covering one opening again. While the apes and the younger children did.

In a series of follow-ups, Redshaw, Suddendorf and colleagues demonstrate that the relationship holds for children from diverse societies (Redshaw et al., 2019) and for a greater range of primate species (Suddendorf et al., 2020). The authors also conducted a socially determined version of the task in which the experimenter dropped the target into one of two parallel tubes, once again children were able to pass, while chimpanzees were not (Suddendorf et al., 2017). Interestingly, only ~30% of 4-year-old children, covered both exits on the first trial, compared to ~80% in the original study, which would suggest that socially determined uncertainty is more challenging for children to comprehend, or that subjects are treating the uncertainty as epistemic rather than physical because the experimenter has already decided which tube they were going to drop the item into, which they find more challenging (Robinson et al, 2006).

1.4 The minimal model of possibility.

Leahy and Carey (2020) developed these two lines of comparative evidence, along with a more extensive range of developmental studies, into their *minimal model of possibility*. They propose that it is only in learning the modal concepts *possible*, *impossible,* and *necessary*, which enables

[8] An adaptation of the forked ramp task (Beck et al., 2006), which was designed to test counterfactuals and future hypotheticals in 4-6 year old children as mentioned above.

children to scaffold a full model of possibility. Prior to this, children only possess a *minimal model* of possibility and, due to the absence of language, any non-human species must also rely on the minimal model.

While a minimal agent can make predictions and simulations as a modal agent can, the minimal agent does not endow their prediction with a symbolic marker to express that it is only a possibility. Instead, they simply make a single simulation and add it to their current model of the world as though it were a well-reasoned belief. Leahy and Carey (2020) explain how a minimal agent could pass a 2-cup task:

> *"Call's cup task is a canonical search task explainable by minimal representations of possibility. Upon seeing the two cups revealed after the hiding event, infants may simulate the prize in one of the cups. If the experimenter then shows that cup to be empty, they simply revise this guess and generate a new one, namely the other cup, which is where they search"(p.72).*

In the 3-cup example, which primates and young children failed, the subject is presented with a target item, hidden in one of two opaque containers (A and B), and a second, equivalent, item hidden in a single container (C). When presented with this choice the minimal agent will make a single prediction as to the state of the world, which contains the location of both items, which they will add to their model of reality. Whereas a modal agent will mark the first piece as "*possibly A, possibly B",* and the second as "*certainly C"*, the minimal agent marks the location of two pieces as *"A*" and "*C*" (or *"B*" and "*C*"). Note that they do not mark the locations as 'certainly', they simply leave them unmarked. They then choose indiscriminately between A and C, because, in the eyes of the minimal agent, both are known. This elegantly explains the 50% of choices directed towards the certain cup by chimpanzees and children under the age of 4. In the forked tube task, a minimal agent only models one outcome, so only covers one exit.

Crucially, the minimal model allows the agent to opt-out or seek more information when they detect cues to uncertainty, features of a task or state which covary with a decreased frequency of success. This means that in metacognitive paradigms such as Call and Carpenter's (2001), they can act as though they are monitoring their own uncertainty while actually lacking the capacity to be uncertain. Specifically, they suggest that running a simulation is an act of last resort. When the agent detects signatures of uncertainty, such as that there is a low contrast between the options or that they hesitated while making their choice, they will then choose to opt out in order to avoid costly time delays.

However, in Hanus and Call's (2014) original task, the $P(1)$ vs $P(0.5)$ was one of a number of conditions featuring a certain outcome. In these trials the uncertain probabilities ranged from $P(0.5)$ to $P(0.167)$. The minimal model predicts that subjects should respond to these in the same manner, as they are choosing between two simulated outcomes of equal strength it should not matter how many cups are in the uncertain array, the subject should choose the certain cup in 50% of trials. However, this is not what the data show, instead the authors report a linear relationship between frequency of taking the certain piece and the probability of correctly choosing the uncertain piece. This conforms with the overall thesis of their paper that instead of appreciating the special value of a certain choice, chimpanzees appear to rigidly rely on the *probability ratio*, the relative likelihood of winning in each gamble.

1.5 Temporal Junctures

In companion piece that appeared alongside Leahy and Carey's (2020) [9], Redshaw and Suddendorf (2020) also tackle the question of how young children think about possibility, they posit that children and primates fail to plan for multiple possibilities because they cannot comprehend the existence of more than one possible timeline. The authors coin the term *temporal junctures* to describe points in time where mutually exclusive timelines diverge. The crucial aspect of the model is that being able to represent temporal junctures means representing that the state of the world is only one of the many which could have been possible at the current point in time. A *live* temporal juncture represents an event which has not happened yet, such as a ball being dropped into a tube. Young children and primates are capable of simulating forwards to catch the ball, but only simulate a single timeline, so they only cover one possible exit. In contrast, past timeline divergences are represented by *expired* temporal junctures, points in time where more than one timeline was possible, but now one has been solidified. While they do not cite Mody and Carey (2016) or Hanus and Call (2014), if an agent only simulates one timeline, then they will choose equally between the certain and the simulated outcome.

Live and expired temporal junctures map onto physical and epistemic uncertainty respectively. However, under the temporal junctures model, epistemic uncertainty requires an additional level of embedding, as the subject must treat the expired temporal junction as if it were live. This requires mental time travel, thought to emerge at around the age of 3-4 years (Suddendorf & Busby, 2005). Under the temporal junctures model, because they must engage in mental time

[9] The authors also pen a joint piece highlighting the complementarity of the two models (Carey et al., 2020)

travel and negate the factual outcome of the world, epistemic uncertainty is conceptually equivalent to counterfactual reasoning, which only emerges around the age of 6 (Beck et al., 2006). As discussed previously, children below the age of 6 effectively prepare for both when facing physical uncertainty, but not epistemic uncertainty (Robinson et al., 2006). To explain the successful performance of younger children in object search tasks, the authors have since argued that below this age subjects do not consider the uncertainty to have been resolved until the cup is lifted (Gautam et al., 2021a), so treat the outcome as though it were undetermined. Redshaw and Suddendorf (2020) suggest that young children reach their conclusion via a Bayesian-like approach, using evidence from previous trials, but once they have settled on a hypothesis, they reject all others. If the two options are represented by approximately equal strengths, then the subject can learn, over the course of an experiment to opt-out in order to avoid punishment in the form of time delays. Like the minimal model, this allows the temporal junctures model to explain why primates appear to show awareness of their uncertainty in metacognition paradigms.

While Redshaw and Suddendorf (2020) are less stringent than the minimal model and permit that non-human animals may be capable of first order temporal reasoning (reasoning about physical uncertainty), they believe that recursion is unique to human (Suddendorf, Redshaw & Bulley, 2022), so suggest that higher levels of temporal reasoning may be outside the realm of non-human species. However, the basis of the temporal junctures model is Hoerl and McCormack's (2017) concept of *event independent time*, which suggests that it is only during development that we learn to think about time as independent from the events that take place. This means representing that an event based label such as "sunrise" on one day is fundamentally different from sunrise on another because of the unidirectional and unique nature of time (McCormack & Hoerl, 2017). Hoerl and McCormack believe in a dual systems approach to temporal reasoning, humans use the *temporal reasoning* system which is reliant on an event independent understanding of time and is unique to humans, while animals and young children possess a *temporal updating* system, which has no abstract representation of time (Hoerl & McCormack, 2019). To represent a temporal junction, one must represent the juncture as a *slot* on a unidirectional linear timeline, which *could* be occupied by more than one event, but *can* only actually be occupied by one. Thus, if they do not think about time in an event independent way, animals cannot reason about temporal junctions.

The temporal junctures model is heavily reliant on the data from the forked-tube paradigm, however, while it is suitably minimalistic, the forked-tube paradigm's utility for assessing future thinking in primates is not without flaws. In follow-up studies, children had no issue with

catching two balls dropped simultaneously into two-tubes (Redshaw et al., 2018; Suddendorf et al., 2020) while the same could not be said for either chimpanzees (Lambert & Osvath, 2018) or monkeys (Suddendorf et al., 2020). Lambert and Osvath (2018), argue that this is due to primates struggling with the coordinated bimanual control to block both exits simultaneously. Bimanual techniques for food processing are rare in chimpanzees and take time to master (Corp & Byrne, 2002) and in their gestural repertoire bimanual gestures are uncommon and basic (Hobaiter & Byrne, 2011). Children's performance on bimanual tasks continues improving in school age children until at least the age of 10 (Schneiberg et al., 2002; Serrien et al., 2014; Serrien & O'Regan, 2021) with evidence that improvement continues between the age of 11 and adulthood (Olivier et al., 2007). While primates and young children can coordinate bimanual action, this is cognitively taxing, so the requirement to plan for mutually exclusive possibilities and synchronised motor control while under a time constraint, may make this task too much of a challenge. Comparative literature has shown repeatedly that additional task constraints can mask cognitive competence in great apes (Seed et al., 2009a). Moreover, when searching for an item dropped into a tube, apes don't reliably choose to search the container under the end of that tube, thus demonstrating a failure to comprehend the tube's causal significance (Cacchione & Call, 2010), which does somewhat lower the overall utility of the task.

1.6 Recent tests of the minimal models in primates.

Due to their declarative and somewhat controversial implications, the minimal- and temporal junctures models created a number of testable predictions which have since been investigated comparatively. Engelmann and colleagues (2021) sought to test whether chimpanzees were able to comprehend epistemic uncertainty using a paradigm which adapted the forked-tube task to test a behaviour which was more familiar to their chimpanzees, rope pulling. The authors presented subjects with two boxes connected by a single rope looped through an eyelet on each box, with a knot between, such that if the subject pulled one end of the rope the knot would catch the eyelet and pull the respective box towards them, but as they dragged the box within reach, the other end of the rope would move out of reach[10]. However, if the subject pulled both ropes simultaneously, both boxes could be dragged within reach. Engelmann and colleagues presented their subjects with two experiments, each of which tested for differences in pulling behaviour between an uncertain and a certain condition but achieved through different means.

[10] This paradigm was originally developed to test cooperation and had the rope-ends placed out of reach of one another, meaning that if the first participant attempted to pull their rope before their partner did, they would lose access to the reward (Melis et al., 2006).

As mentioned, chimpanzees do not naturally perform well in tasks which require bimanual coordination, so simultaneously pulling both ropes represents a cost, so was the dependent measure.

In the first experiment the boxes were either opaque or transparent, and the subjects only pulled on both ends of the rope in the opaque condition. In experiment 2, conducted with a different group of chimpanzees, subjects were presented with two opaque boxes, in the certain condition they were allowed to observe the baiting from an adjacent room, whereas in the uncertain condition they were not. As in experiment 1, subjects only pulled on both ropes in the uncertain condition. The authors take these results as a reflection of apes being able comprehend epistemic uncertainty and, therefore, they contradict the minimal model. However, an alternative interpretation, is that the chimpanzees were simply seeking information by pulling both in, therefore delaying the point of decision and the necessary simulation until they had both boxes within reach, this second explanation does not conflict with the minimal model. A point worth noting is that, in cases where a chimpanzee pulled both ropes and obtained both boxes, they were significantly less likely to open the second box if they had found the apple in the first. This suggests that firstly they have an exclusive understanding of the disjunction (A *or* B *not* both) but secondly that they were confident enough to not open a second box which was within arm's reach.

Ferrigno, Huang and Cantlon (2021) attempted to make the 4-cup task more closely resemble natural foraging by removing the second experimenter, whose role it was to make the first incorrect guess, and instead allowing the subject to take two guesses to find the food items. Their logic being that, provided the subject could not detect the location of the food, they should locate a food item on half of their first guesses. Their subjects were 9 zoo-housed olive baboons (*Papio anubis*), of which 4 passed on to the test phase. When they failed to find a food piece on their first guess, the 4 baboons switched pairs on 41% of trials, significantly lower than the 66% chance predicted by numerical chance. At an individual level, 3 of 4 subjects switched at rates significantly below chance and 2 of these individuals did so below the 50% level expected by the minimal model. However, subjects did receive a large number of trials (185-229), and the data did show a small learning effect. Based on their statistical model, subjects still would still have passed from the first trial but nevertheless it would be valuable to have performance data from the first session. In comparison, subjects performed almost precisely at chance in the reveal baited condition, switching on 66.3% of trials. The authors don't report individual data for reveal baited trials, which would suggest that none of the subjects were individually above chance.

However, I propose that sequential search naturally lends itself to passing reveal empty trials. Firstly, in sequential search paradigms primates have a documented difficulty with inhibiting searching of adjacent cups (Barth & Call, 2006), a strategy of always searching the adjacent cup would be sufficient to pass reveal empty trials at a 33% chance rate. If we first assume that when presented with the 4-cup array on their first guess a subject chooses equally between the inner and outer cups, then a policy of searching the adjacent cup will result in success on 100% of all outer guesses, where the only adjacent cup is the pair, and 50% of inner guesses, where there is a cup either side. It would be valuable to have access to trial-by-trial data, as the learning effect could have been a trend towards choosing the outer cups. Secondly, reveal empty trials would be a better setting for this strategy because they mimic searching for a single piece, while on reveal baited trials, having already found a piece, it is possible that the subject then approached the guess as a new search, and chose randomly between the 3 remaining cups.

Gautam, Suddendorf and Redshaw (2021a) argue that the baboons were only reasoning via the inclusive disjunction, either A *or* B, *not* A *therefore* B and were precisely at chance in the exclusive disjunction, A *or* B, *not* A *and* B. This, they argue, means that the baboons do not represent *or* as a true logical operator. This would reflect that the baboons are using abductive inference, that A or B *might* contain the grape. Concurrently the authors were running a similar study with children where they repeated Mody and Carey's (2016) 4-cup task but modified it to include *remove-sticker* trials, in which the sock puppet who was taking the first guess successfully found the sticker (Gautam et al., 2021b). They found that while children as young as 2.5-years-old were above chance in the remove-empty condition, it was not until age 5 that children scored above chance in the remove-sticker condition. They suggest two interpretations, equally extensible to baboons, either that young children exclusively treat the *or* relation as inclusive, that the sticker *must* be under A or B but failing to represent that it cannot be under *both* A and B; or alternatively that they are able to make affirmative inferences before their comparable negations, if A is empty then it must be under B, but not that if A was full it must not be under B.

In a wide ranging study which tested chimpanzees under the 2-, 3- and original 4-cup paradigms, Engelmann and colleagues (2023) confirmed previous findings from the great ape literature, subjects answered the correctly on 95% of 2-cup trials but only 51% of 3-cup trials. Finally, in the original 4-cup paradigm, which chimpanzees had not previously been tested under, the authors found that subjects were at chance in the reveal empty condition, switching pairs in 48% of trials, but performed close to ceiling in the reveal baited condition, switching on 85% of trials.

The authors propose an alternative non-cognitive *location-based* explanation as to how chimpanzees may be solving this task. Under this explanation, an ape marks a broad location which comprises all of the cups a target item could be under, if they receive confirmatory evidence about the contents of a cup, visually or otherwise, they shrink the location to only include that cup; if they receive negative evidence then they shrink the location to exclude that cup, and if they see a piece taken away they remove that whole location. When allowed a choice, they pick indiscriminately between any of the broad locations which remain, without appreciating the number of individual cups which it covers. In the 3-cup task the subject marks two locations one covering the uncertain pair and the other covering the certain cup, because they receive no further information, they pick indiscriminately between the two broad locations, resulting in the 50% choice rates observed. In the reveal empty condition of the 4-cup task the outcome is fundamentally the same, but the subject must reach the conclusion that the certain cup is certain by inference. In reveal baited trials, the subject sees one food piece removed, so removes that whole location, leaving only the broad location covering the uncertain pair, which they search, resulting in the near ceiling performance.

The location based theory also holds for the original 2-cup task and Call's revised 3-cup-1-item task (2022). In the 2-cup task there is only ever 1 location, when it is shrunk by new information it now only contains one cup, so the ape chooses it correctly. In the 3-cup-1-item task there is one location covering the possibly baited cups. In the reveal empty condition, the subject shrinks it to only cover the remaining cup, so chooses correctly, and in the reveal baited condition they see the piece removed so remove the entire location and pick indiscriminately between two empty cups. However, this explanation does not hold for baboons (Ferrigno et al., 2021) or children (Gautam et al., 2021b) and the authors note that it does not explain why the apes tested in Hanus and Call's (2014) original study passed at the 1 vs 1/6 condition.

1.7 A probabilistic framework

Rescorla (2009) introduces a probabilistic framework to explain the actions of Chrysippus' dog. Rather than approach the problem deductively, he argues that we only need to ascribe Bayesian inference to our hypothetical dog. If we first assume that the dog starts with an equal probability of its quarry having fled down each of the 3 paths and that there is no chance it has gone elsewhere, there is a $\frac{1}{3}$ likelihood that it went down the first path and a $\frac{2}{3}$ likelihood that it did not, therefore the dog should sniff the path to check. Having received negative evidence from the first path it then updates the likelihood of it having gone down the remaining paths to be 1-x where x is the likelihood of a false negative from path 1. As it approaches the second path, the

likelihood of being correct is $\frac{1-x}{2}$ while the chance of being incorrect $\frac{1-x}{2} + x$ is higher[11], so the dog should seek more information. After receiving more disconfirming evidence from path 2, the dog updates its priors again about the 3rd path being correct to (1-2*x*) and being incorrect to 2*x*. Provided that the likelihood of a false positive is not greater than 25%, then the dog can simply go down the third path without sniffing.

To extend this probabilistic account to the primate literature reviewed here, in the metacognition tube task (Call and Carpenter, 2001) subjects performed the same inference that was performed by the dog, but also showed an increased likelihood to check after a time delay. Increasing the time delay between receiving the information and acting on it increases the likelihood of a false negative, thus increasing the value of a final check, which the data support. While adding in a measure of resource quality, and thus potential loss, can explain why, even in the visual condition, apes are more likely to check even when they have observed the baiting if the value of the item is high (Call, 2010). Notably, for a dog who is under time pressure pursuing an animal, stopping to sniff may represent a high cost, while for an ape, checking the tube does not. Plausibly, re-running the metacognition task with an increased cost to checking such as through some form of time pressure, be it from a competitor or a diminishing resource, we may see a higher reliance on inference.

Rescorla (2009) notes that the mathematical literature is filled with instances of calculations in the absence of mathematics, citing Turing machines as a classic example. As described above, Friston (2010; 2009) has continually argued that even the simplest organisms are predictive engines that use Bayesian inference to minimise 'surprise' – the difference between what is predicted and what is experienced. The question undoubtedly remains however, whether these animals are aware of their uncertainty, that is whether they have an intuitive strength of representation without being able to consciously represent its strength. The probabilistic account ascribes broadly rational behaviour to animals, so it does not predict the poor performance of chimpanzees in the 3-cup task and reveal empty trials of the 4-cup task, meaning that if it were broadly correct then there must be additional constraints limiting the apes' performance.

1.8 Ratio of Ratios

When discussing their original results Hanus and Call (2014) reached a similar conclusion to Rescorla (2009), arguing that chimpanzee's reason probabilistically rather than

---

[11] $\frac{\text{Total} - p(\text{false negative from path 1})}{\text{Number of paths remaining}}$ vs $\frac{\text{Total} - p(\text{false negative from path 1})}{\text{Number of paths remaining}} + p(\text{false negative from path 1})$

deterministically, and believe that apes do not ascribe special status to a certain outcome as humans do. They propose that, instead, apes treat both outcomes as uncertain choices with different magnitudes of uncertainty, after which they rigidly follow a ratio-of-ratios (RoR) approach. The RoR approach can be thought of as measuring the chance of success offered by each option and then assessing their size relative to one another (Eckert, Call, et al., 2018). For example, choosing between a $\frac{3}{5}$ chance and $\frac{1}{4}$ chance would be $\frac{1.5}{0.25}$, which an RoR of 6.

This approach is a hallmark of the analogue magnitude system (AMS), a quantity estimation system conserved in humans, primates, and other animals (Cantlon et al., 2015). The AMS compares the relative sizes of neural signals representing external stimuli, with discriminability following a logarithmic curve. Small relative differences are indistinguishable, but above a certain threshold, performance is close to ceiling. Crucially, because the system relies on relative sizes rather than absolute values, the size of the samples it can estimate are unlimited, however the error scales with size, meaning that the smallest ratio which can be discriminated remains constant, this is known as the Weber fraction.

If the Weber fraction of their uncertainty resolution system is greater than 2:1 then the apes are unable to discriminate between the certain and the uncertain choice in the 3-cup task, so choose randomly between the two options. While in more unequal comparisons (such as 4:1 and 6:1), subjects would be able to resolve the uncertainty, so would choose adaptively. This can also explain the performance of apes in both variants of the 4-cup task, but, like the location-based argument, not the 4-cup performance of baboons or children.

A RoR approach also finds support in the statistical inference literature, in which infants and great apes reliably choose which of two samples, each containing desired and less desired items, is more likely to give them a desirable outcome (Denison & Xu, 2010, 2014; Eckert, Call, et al., 2018a, 2018b; Rakoczy et al., 2014; Xu & Garcia, 2008). In great apes and, somewhat surprisingly, also in adult humans, performance breaks down at RoRs between 2 and 4 (Eckert, Call, et al., 2018), as it did in the 3-cup task. Significantly, both apes and infants are able to take into consideration the experimenter's preference but revert back to using statistical inference when the experimenter is blindfolded (Eckert, Rakoczy, et al., 2018). Which suggests that these groups are specifically measuring their magnitude of uncertainty, rather than simply discriminating between the uncertainty of the populations.

Notably, the RoR approach can explain the final stage of the location argument. If the subject is presented with a location containing 1 item and 1 cup (1/1 chance of a positive outcome) and a location containing 1 item and 2 cups (½ chance), they fail to discriminate the RoR of 2. While

in the 1 vs 1/6 array the RoR is 6, which the apes can discriminate. However, this is still assuming that apes are selecting between 'risky' and 'safe', which is more applicable to the urn task. If they were instead representing the likelihood of each cup containing a grape, then if they fail to discriminate p = .5 from p = 1, then they would instead choose the uncertain outcome 66% of the time rather than 50%. If we were to specify instead that the subjects were focussing on the food items not the cup, and were reflecting on their likelihood of success then these arguments can work together to explain performance.

A failure of uncertainty discrimination would make predictions for decision making under risk and uncertainty. Apes are adept at discriminating reward size (Schmitt et al, 2013[12]) and quantities (Hanus and Call 2007), so discriminating value but not uncertainty would manifest in being irrationally risk prone.

In object choice studies, where the odds are visible, we do find support for this. An applicable study used a cup-based search task to test the risk profile of great apes as the experimenters systematically varied the likelihood of success and the relative sizes of the safe and risky pieces (Haun et al., 2011). The apes were presented with a constant safe choice, and a choice of 1-4 cups, one of which contained a food piece 1.5, 3 or 6 times larger than the safe choice, resulting in expected values of the risky choice ranging from 6 to .375. In line with the RoR account, apes only took the safe piece in the 1.5x condition, when EV of the risky choice ranged from 1.5 - 0.375, and didn't differentiate between the number of risky cups [13]. In the 1.5x condition apes chose the safe piece 50% of the time, so feasibly this could have been a failure to also discriminate the size difference and resulting in choosing equally between the two. While, apes did not show this effect in visible trials, this could be a consequence of additive errors within the analogue magnitude system when estimating size and uncertainty, as have been proposed by Eckert et al. (2018).

In an urn-like task with only 2 outcomes per urn, apes were approximately indifferent between a certain option, where both outcomes were a single peanut, and an uncertain option where the outcome was two peanuts or nothing, which is behaviour in line with expected value and not the

---

[12] Although gorillas were indifferent in the small size discrimination condition in this task (cubes with a side length of 44mm vs 50mm side-length), they performed better in a large format version and, in the study discussed subsequently, they correctly selected the larger piece on ~90% of visible trials at the smallest size difference (Haun et al., 2011).

[13] While the authors do not conduct a separate analysis for each level of reward size, they failed to find an overall effect of risky cup number in hidden trials, and the median rate of taking the safe piece for the 1, 2, 3 and 4 cup conditions of the 1.5x trial type were 0.75, 0.5, 0.75 and 0.5 respectively.

uncertainty resolution hypothesis (Haux et al., 2023). In monkeys the evidence is mixed, in a 4-cup gamble with equal expected value between the risky and safe rewards, capuchins *(Sapajus apella)* have been reported as being indifferent (Rivière et al., 2019) or highly risk averse (Roig et al., 2022), while mangabeys (*Cercocebus torquatus)* were highly risk prone (Rivière et al., 2018).

A greater number of studies of uncertainty in primates have employed a paradigm analogous to variable reinforcement, in which subjects are presented with a safe and a risky bowl, the safe reward contains a constant intermediate reward, while the contents of the risky bowl could be comparatively better or worse. Within the *Pan* lineage, studies using these paradigms have shown reliable divergent risk profiles, with chimpanzees being more risk prone than bonobos in both quantitative (Heilbronner et al., 2008; Keupp et al., 2021) and qualitative variants (Rosati & Hare, 2010, 2012, 2013). This is an essential component of a broader explanation in the risk literature, which suggests that chimpanzees and orangutans are more risk prone because they have a higher proportion of fruit in their diet, which is calorie dense but sparsely dispersed, while bonobos and gorillas rely on readily available herbs and browse, which is reflected in their more conservative risk profile. Nevertheless, as the AMS system is evolutionarily ancient, if it was governing choices under variable reinforcement, we would see the same response from bonobos and chimpanzees.

This would suggest that RoR system is applicable to tasks where the odds of success are visible, but not ones where the frequency of success is learned through experience. Another possibility could be that apes use individual preference to choose between a variable vs fixed reward when recalled from memory, but the RoR is used to distinguish between different variable reward schedules. Steelhandt et al. (2011) showed that around half of monkeys distinguished between a cup that gave 9 raisins on 2/3 of trials and one that gave 18 raisins on 1/3 of trials, an RoR of just 2, so this cannot be the case. This evidence suggests that extending the RoR approach as a more general explanation for how great apes view uncertainty is a stretch. Notably, the location-based argument, the minimal model and the temporal junctures model all also cannot explain choice behaviour in line with expected value.

1.9 The current work.

Following this general overview of reasoning in great apes, this thesis aims to test some of the explicit claims laid out by the varying models of non-linguistic reasoning. I present four experimental chapters followed by a chapter discussing how my results link to the wider literature.

In Chapter 2, I set out to explicitly test Leahy and Carey's (2020) minimal model of possibility. In experiment 1 I test the sequential guessing strategy which the authors lay out for how a minimal agent would conduct a directed search. To do so I developed a novel repeated choice paradigm I term *post-decision wagering*, whereby I first gave the subject a choice in a simple 2-cup 1-item task while manipulating whether they have visual access to the baiting. After they had made their choice but before revealing they were correct, I offered them the choice between their original guess and a fractional reward. If, as predicted by the minimal model, chimpanzees only revise their predictions in light of new evidence, then we should see no difference between trials where they observed the baiting and trials where they did not. The data do not support this conclusion so we can reject this strict version of the minimal model. In Experiment 2 I repeated this paradigm with all 4 great ape species and with a larger, more diverse cohort. The data shows that there was no difference between the species, but that there was a quadratic relation between age and confidence, with younger and older individuals taking the half-grape more in uncertain trials. In experiment 3 I repeated the paradigm but controlled for the strength of the representation by making both trial-types occluded, but manipulated how many cups were behind the barrier during baiting. Once again, we found the same results, meaning that we can reject the minimal model of possibility as an explanation of great ape choice behaviour.

In Chapter 3 I deploy the same post-decision wagering paradigm but give the subjects information about their unchosen cup before the second choice and see whether they are able to adaptively alter their choice behaviour. This involves reasoning via the disjunctive syllogism and cannot be solved by simply avoiding the empty cup, but without the working memory constraints of tracking 2 food items in 4 cups. In Experiment 1 I find that, while as a group, chimpanzees did not adaptively alter their half-choice rates, 2 individuals did do so, and one scored 100% on both reveal-empty and reveal-baited trials. This suggests that he was able to reason via both the inclusive and the exclusive disjunction, thus this behaviour is not unique to humans. In Experiment 2 I once again extended this research to a more diverse cohort, finding that there was no difference between the species, but that the individual base-rate of taking the half-piece in Chapter 1 did predict overall performance in this experiment. Once again, I found one individual who was near ceiling in both conditions, which shows that the individual in Experiment 1 is not unique in this ability. In Experiment 3 I return to the original chimpanzee group with a revised paradigm which controlled for the individual having solved the task by stimulus enhancement. In the revised task the group were above chance in both conditions, and 6 of 9 individuals altered their half-choice rates adaptively. Notably, the individual who scored 100% on the previous task scored 96% on the revised task. I propose an alternative, inhibition related, explanation for why apes are able to pass 2-cup tasks but fail at the 3- and 4-cup task.

In Chapter 4 I test all 4 great ape species on the Mody and Carey (2016) 4-cup task and the Ferrigno et al., (2021) modification. I find that in both variants apes adaptively switch between pairs in response to the contents of the revealed cup, but that performance is better in reveal empty than reveal baited trials which reinforces the results of Ferrigno and Gautam and a continuation between monkeys, apes and humans. Four subjects switched adaptively between the pairs in response to the revealed cup contents, all in the Ferrigno modification. One individual was above chance in both the reveal empty, and the reveal baited condition which is unique within the literature. In Experiment 2 I retested those individuals who had passed the original experiment but included control trials in which both food pieces were placed into 1 pair of cups, this was to test an associative strategy of '*win-switch lose-stay*'. Crucially, this tested whether subjects were able to flexibly apply the disjunctive syllogism when required. I compared the performance of this group to a naïve group of chimpanzees and found no difference in performance, which suggests that the original group have not been conditioned into this response. However, overall performance was low, which suggests that in the earlier experiment apes were not solving the disjunction logically.

In Chapter 5, I test whether chimpanzees are curious about counterfactuals. I utilised Call and Carpenter's (2001) 3-tube paradigm to test whether, after having received the contents of their choice, chimpanzees would expend effort to check what was in the tube they did not pick. I did so by covertly baiting 3 covered tubes with either a large piece of apple, a small piece of apple, or nothing, then giving the subject the choice between 2 of the 3. Meaning that knowing what they did get does not tell them what they passed over. After receiving their choice, I uncovered the ends of the unchosen and unavailable tubes and recorded which tubes they checked. I found that subjects checked the available but unchosen tube more often than the unavailable and also checked the unchosen tube first. This showed that they were curious about the counterfactual rather than simply resolving uncertainty. In Experiment 2 I reverted to a simple metacognition paradigm to test a reductive explanation that they simply checked the tube which was closer to them (the unchosen tube), however the data do not support this conclusion. In Experiment 3 I investigated the importance that agency has in chimpanzees' counterfactual curiosity, finding that, although it does not impact the absolute frequency with which they search for counterfactual information, subjects only bias their search towards the unchosen tube when they had agency over the choice.

Finally, in Chapter 6, I discuss how the experimental evidence presented in this thesis links to one another and how it relates to the evidence presented in this chapter. This discussion will

emphasise the relevance of the studies presented here and how they can advance our understanding of how great apes view possibility.

2. Differentiating possibility from certainty.

The data presented in experiments 1 and 3 of this chapter have been published as part of the following paper:

Jones B, Call J. Chimpanzees (Pan troglodytes) recognize that their guesses could be wrong and can pass a two-cup disjunctive syllogism task. Biol Lett. 2024 Jun;20(6):20240051. doi: 10.1098/rsbl.2024.0051.

2.1 Abstract

When great apes search for hidden food, do they realise that their guesses may not be correct? We applied a post-decision wagering paradigm to a simple 2-cup search task, varying whether we gave participants visual access to the baiting and then asking after they had chosen one of the cups whether they would prefer a smaller but certain reward instead of their original choice (experiment 1). Results showed that chimpanzees were more likely to accept the smaller reward in occluded than visible conditions. In experiment 2 we extended this result to a more diverse cohort comprising all 4 great ape species', who demonstrated a consistent effect across species but a quadratic relation between age and confidence in uncertain trials. Experiment 3 returned to the original cohort and found the same effect when we blocked visual access but manipulated the number of hiding locations for the food piece, showing that the effect is not due to representation type or non-cognitive clues to uncertainty. Overall, these results show that great apes do not treat their guesses as equivalent to well-reasoned beliefs and can comprehend the existence of multiple incompatible possibilities.

2.2 Introduction

Search tasks are a valuable tool in in comparative psychology as they mimic natural foraging. Rational choice behaviour in these tasks can be considered evidence of logical thought. Notably, evidence in non-human primates (Engelmann, Haux, et al., 2023; Hanus & Call, 2014) and young children (Gautam et al., 2021b; Mody & Carey, 2016) suggests that these groups fail to appreciate the unique value of a certain outcome, choosing it approximately equally to an alternative with P = 0.50.

This has led some authors to argue that, due to lacking the language of modal concepts, animals and preverbal infants lack a full model of possibility and instead use a minimal model reliant on making a single simulation of reality, which they act on without considering alternative possibilities (Leahy & Carey, 2020). With the aid of a hypothetical example the authors describe a process of *sequential guessing* (Box 1, p67), whereby a minimal agent, in this case a young chimpanzee searching for its mother, will use past frequencies to make a prediction of the actual state of the world, including the actual location of its mother, which they will then add to their model as though it was certain knowledge. If, however, upon searching, that prediction turns out to be incorrect, they will then make a revised prediction, which they act on again. While this is sufficient for search in the wild, in the experimental setting this manifests as choosing indiscriminately between the actual- and the simulated certain outcomes, because in the eyes of the minimal agent, both are known.

The recent evidence testing the minimal model of possibility in non-human primates has been mixed. In a 1-item 2-location search task, chimpanzees acted to maintain access to both locations while they searched, but only if they had not observed the baiting (Engelmann et al., 2021), behaviour which conflicts with the minimal model. However, in a repetition of Hanus and Call's (2014) 3-cup 2-item task, the same group of chimpanzees failed to choose the certain cup above chance rates (Engelmann, Haux, et al., 2023), thus the authors note that they cannot reject Leahy and Carey's hypothesis. This inconsistency could be explained as either that chimpanzees struggle with the working memory constraints of tracking 2 items, or that maintaining access to both search locations is a form of information seeking and doesn't require simulating either scenario until the point of searching.

This second argument is in line with a more refined version of the minimal model laid out by Leahy and Carey (2020). The authors describe how, when faced with a task in which errors are costly, a minimal agent can act as though they are monitoring uncertainty without being aware of her uncertainty. They do so by learning to recognise perceptual features of a presentation which have been previously associated with a decreased frequency of success and choosing to opt-out or seek more information in these situations.

In the current study we first tested 9 zoo-housed chimpanzees in a novel repeated-choice paradigm that retrospectively probed participants' certainty in their answers in the absence of new information. We used a standard 2-cup 1-item search task but manipulated whether subjects had visual access to the baiting procedure. After the subject had chosen but without revealing whether they were correct, we offered them the choice between their original selection and a visible fractional reward. If the subjects were behaving in line with the sequential guessing

hypothesis, then we would see no difference in rates of taking the fractional piece between the visible and occluded condition. In Experiment 2, we extend this research to a more diverse group which included members of all 4 great ape species and a larger range of ages. Finally, in Experiment 3, we adapted the paradigm to test for the more refined version of the minimal model of possibility, by matching the modality of the presentation.

2.3 Experiment 1

Methods

Participants.

We tested 9 chimpanzees aged between 7 and 46 (3 female, mean age = 31.2 years), full demographic details can be found in Table 2.3.1. Subjects were housed at the Budongo Research Unit (BRU), which operates within Royal Zoological Society of Scotland's Edinburgh Zoo. The subjects live in a natural group, enclosures allow access to both indoor and outdoor space with vegetation. The chimpanzees receive regular feedings throughout the day which are comprised of a wide variety of fruits and vegetables, the group additionally receives further enrichment. Individuals are experienced in non-invasive cognitive testing and similar search paradigms. Testing is voluntary, non-contact and takes place in a communal area accessible to all group-members, at no point were subjects separated from their group.

Table 2.3.1: Demographic details of the participants of Experiment 1 and 3.

ID	Sex	Rearing History	Age	
			Exp 1	Exp 3
Edith	F	Parent	25	27
Eva	F	Parent	41	42
Frek	M	Parent	28	29
Kilimi	F	Parent	29	30
Liberius	M	Parent	-	24
Louis	M	Wild-caught	45	-
Lucy	F	Parent	45	46
Masindi	F	Parent	-	3
Paul	M	Parent	28	30
Qafzeh	M	Parent	29	31
Velu	M	Parent	7	9

Materials.

All demonstrations took place on a sliding table (630mm x 300mm) attached to the outside of the enclosure, in its forward position the table pressed against the plexiglass panel and subjects

were able to indicate their choices by placing a finger into one of three 'choice holes' at the base of the plexiglass panel. Two identical plastic cups (Ø = 86mm, height = 89mm) were used as hiding locations, they were placed in front of the outer two choice holes, 10cm from the front of the sliding table and 10cm in from either edge. An occluder (height = 250mm, width = 450mm) was used to block the subject's view of the cups during occluded trials. Whole grapes were used as the target item, and halved lengthways for the fractional reward.

Procedure

Training trials.

As subjects were familiar with cup-based search tasks due to previous research, comprehension trials focused on subjects being aware that they could choose a half grape instead of a cup. Stage one consisted of the experimenter placing one cup face down on the table empty, then a half grape on the opposite side of the table before sliding the table to the participant for them to indicate their choice, passing this stage required that the participant choose the half grape on two consecutive trials. Stage two involved the experimenter placing two empty cups down onto the table, sliding one back and then placing the half grape in its place and offering the choice to the subject. Passing this stage also required that the subject choose the half grape on two consecutive trials. In stage three, the experimenter placed one grape onto the table, then covered it with a cup, then placed an empty cup on the other side of the table, the empty cup was then slid back and replaced with a half grape, passing this stage required that participant ignore the half grape and select the baited cup. In the second trial of this stage, the first cup was left empty while the second cup was baited, the empty cup was again slid back and replaced by a half grape. If subjects failed any two trials consecutively then the experimenter abandoned testing for that session, and the subject started from comprehension stage one at the next session.

Test Trials

Test trials were divided into *visible* and *occluded* conditions. In *occluded* trials, the experimenter first lifted both cups to show that they were empty then placed the occluder between themself and the subject. They then held one grape above the occluder and ensured that the subject's attention was drawn to it, they then brought it down behind the occluder at its centre. The experimenter visited the first cup and lifted it with the corresponding hand (left hand lifted left cup), before bringing both hands together in the centre and repeating with the second cup depositing the grape under one of them. The experimenter then lifted the occluder and placed it on the floor before sliding the table to the subject to allow them to indicate their choice via the choice holes. Once they had chosen, the experimenter pulled the table back to themselves and moved the unchosen cup to the back of the table without lifting it while also

touching the chosen cup. They then placed a half grape in place of the removed cup and slid the table back to the subject to make a second choice. *Visible* trials were identical to occluded trials but without the occluder, allowing the subject to observe the baiting. Subjects received 12 visible and 12 occluded trials, each 12-trial block was arranged pseudo-randomly and contained 6 visible trials and 6 occluded trials and were counterbalanced by order in which the cups were visited.

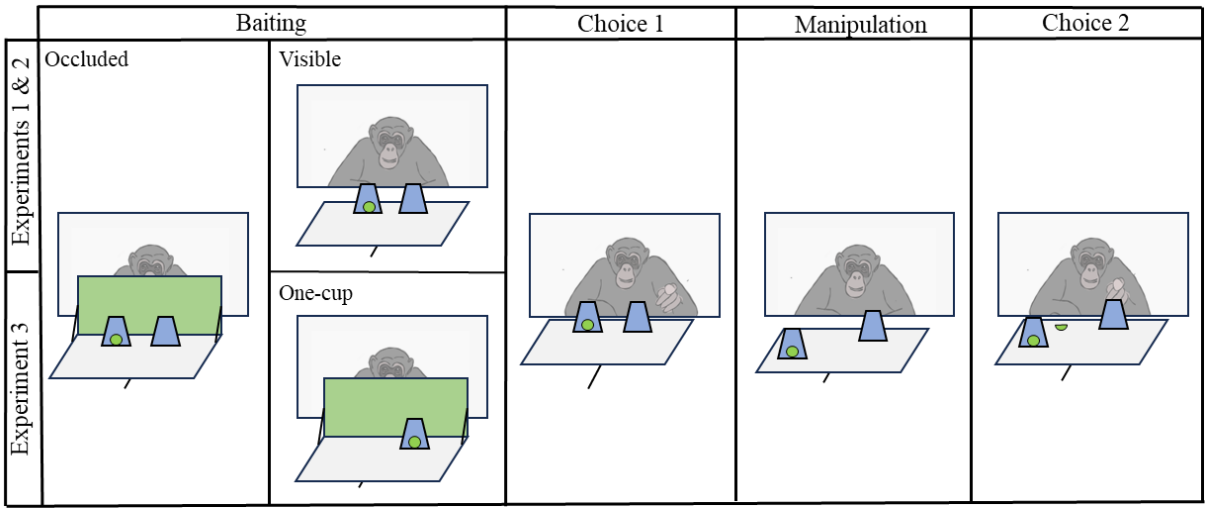


Figure 2.3.1: Procedure for test trials in experiments 1-3. Adapted from Jones and Call (2024)

Data Coding and Analysis

The experimenter live coded first and second choices and 15% of trials were recoded from video recordings by a second experimenter blind to the purpose of the experiment. Inter-coder reliability based on the first and second choices for 15% of trials was perfect (kappa = 1, n = 59). Data analysis was conducted in R (version 2021.09.1).

Results and Discussion.

Figure 2.3.2 presents the percent of trials in which subjects chose the half grape as function of condition. Subjects chose the half grape significantly more often on occluded trials than on visible trials (t-test, $t = 5.08$, $df = 8$, $p < .001$). Demonstrating that chimpanzees are not equating a guess with a certain outcome. Under the minimal model, a minimal agent will "use simulation to generate a single result and treat that result as reality" (Leahy & Carey, 2020, p. 67), therefore the chimpanzees tested here are not conforming to the predictions of the minimal model.

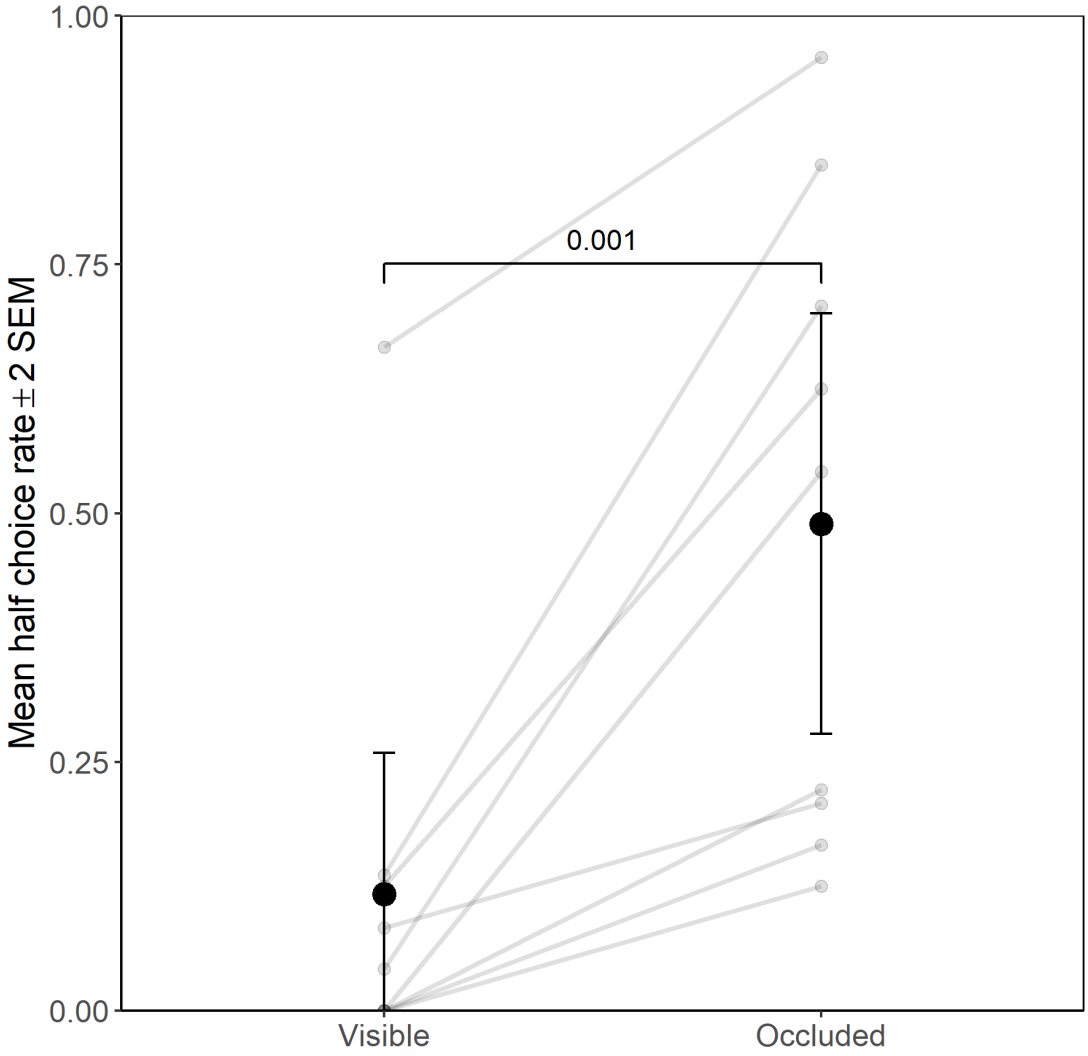


Figure 2.3.2 Group level differences in experiment 1. Error bars show 2 standard errors around the mean, lines and small points show individual level means.

There was a large amount of inter-individual variation in confidence (Figure 2.3.2, Table 2.3.2 Individual rates of taking the half grape in experiment 1 by condition. Fisher's exact test (one-sided) for contingencies between condition and half-choice rate and adjusted for multiple comparisons (Holm-Bonferroni).). However, all individuals chose the half grape more frequently in the occluded condition than in the visible and this difference reached significance for 2 individuals (Fisher's exact test, one-tailed, p = .025).

Table 2.3.2 Individual rates of taking the half grape in experiment 1 by condition. Fisher's exact test (one-sided) for contingencies between condition and half-choice rate and adjusted for multiple comparisons (Holm-Bonferroni).

ID	Visible	Occluded	Fisher Test (p)	Fisher Test (p.adj)
Edith	0	0.222	0.041	0.075
Eva	0.083	0.208	0.208	0.208
Frek	0.042	0.708	< .001	< .001
Kilimi	0	0.542	< .001	< .001
Louis	0.667	0.958	0.011	0.025
Lucy	0	0.125	0.117	0.132
Paul	0.125	0.625	0.059	0.076
Qafzeh	0.136	0.85	0.163	0.435
Velu	0	0.167	0.055	0.076

Redshaw and Suddendorf (2020) argue that non-human primates and young children can act adaptively in uncertain situations without being aware of their uncertainty. They point to classic uncertainty monitoring paradigms and suggest that subjects learn through time punishments to opt-out of a decision when representations are approximately equally weighted, and failure to opt-out of difficult trials is shaped by operant conditioning. In our experiment, there was no time penalties and post-decision wagering asks subjects to rate their decision retrospectively so there is no option to opt-out before trials. We found no difference in rates of taking the half grape between sessions, and while chimpanzees have not engaged with either traditional uncertainty monitoring paradigms or experiments involving time delays, it is possible that that their extensive prior experience may have offered some carryover. In Experiment 2 we test this conclusion by repeating the procedure with a naïve group while also extending the research to the other great ape species.

2.4 Experiment 2.

Methods

Participants

We tested 25 zoo-housed great apes (10 bonobos (*Pan paniscus*), 6 chimpanzees (*Pan troglodytes*), 5 gorillas (*Gorilla gorilla*) and 4 orangutans (*Pongo pygmaeus*), Table 2.4.2), ranging in age from 3 to 57 (mean = 22.8). The subjects were housed at Twycross Zoo, England and live in species typical groups with access to both indoor and outdoor spaces with vegetation. They receive regular produce feedings of a variety of vegetables throughout the day alongside additional enrichment and routine training. The majority of individuals were born in

captivity, with the exception of three individuals who were wild-caught (*Coco, Samantha, Biddy*) and one whose provenance is unknown (*Likemba*). Subjects had engaged in intermittent cognitive testing in the preceding 12-months, for the gorillas and orangutans this was their first experience with object search tasks while the bonobos and chimpanzees had participated in one previous study.

Table 2.4.1 Demographic details of the participants of experiment 2.

Species	ID	Sex	Age	Rearing History
Bonobo	Cheka	Female	26	Parent
	Daitou	Female	44	Hand
	Likemba	Female	12	Unknown
	Lina	Female	37	Parent
	Lola	Female	3	Parent
	Lopori	Female	10	Hand
	Lucuma	Male	19	Parent
	Malaika	Female	12	Parent
	Ndeko	Male	7	Parent
	Rubani	Male	5	Unknown
Chimpanzee	Coco	Female	57	Wild-caught
	Holly	Female	39	Parent
	Josie	Female	34	Hand
	Kibali	Male	18	Parent
	Samantha	Female	42	Wild-caught
	Tuli	Female	15	Parent
Gorilla	Biddy	Female	48	Wild-caught
	Lope	Male	9	Parent
	Oumbi	Male	30	Parent
	Ozala	Female	28	Parent
	Shufai	Male	5	Parent
Orangutan	Basuki	Male	5	Parent
	Batu	Male	33	Parent
	Kayan	Female	5	Parent
	Maliku	Female	28	Parent

Apparatus.

The experimenter sat opposite the subject with a sliding table between them. The table was positioned so that when it was in its forwards position the subject could reach their fingers through the mesh to select a cup by touching it. The table, cups and occluder were the same as experiment 1. In place of grapes the concealed food was pieces of raw sweet potato, the large piece was twice the volume of the fractional piece, $2cm^3$ and $1cm^3$ respectively. For one

individual, Batu, 2 fractional pieces were used in place of the large reward due to him not reliably selecting the large piece during the initial training phase.

Procedure.

The procedure for the pre-test and test trials was identical to experiment 1, two additional apes failed or did not complete the pre-test.

Data Coding and Analysis

The experimenter live coded first and second choices, inter-coder reliability based on the first choice, removed cup contents and second choice for 15% of trials was excellent (kappa = .954, n = 112).

Results and Discussion

Figure 2.4.1 presents the percent of trials in which subjects chose the half fractional piece as function of condition. Subjects chose the half piece significantly more often on visible trials than on occluded trials (t-test, $t_{24} = -5.57$, $p < .001$), thus replicating the data from the Edinburgh group and, providing evidence against this strict version of the minimal model of possibility . Furthermore, we find no difference in the rate of taking the half piece between the experimentally experienced Edinburgh group, and the naïve Twycross group (Visible: Wilcoxon test, W = 137.5, 100.5 p = .643. Occluded: Wilcoxon test, W = 137.5, p = .338) countering the suggestion that the Edinburgh results could have been influenced by their experience. Table 2.4.2 shows individual rates of taking the half piece by condition. After correcting for multiple comparisons, we find that 7 individuals took the half piece more frequently in the occluded than the visible condition (Fisher's exact test, one-tailed, p = .05).

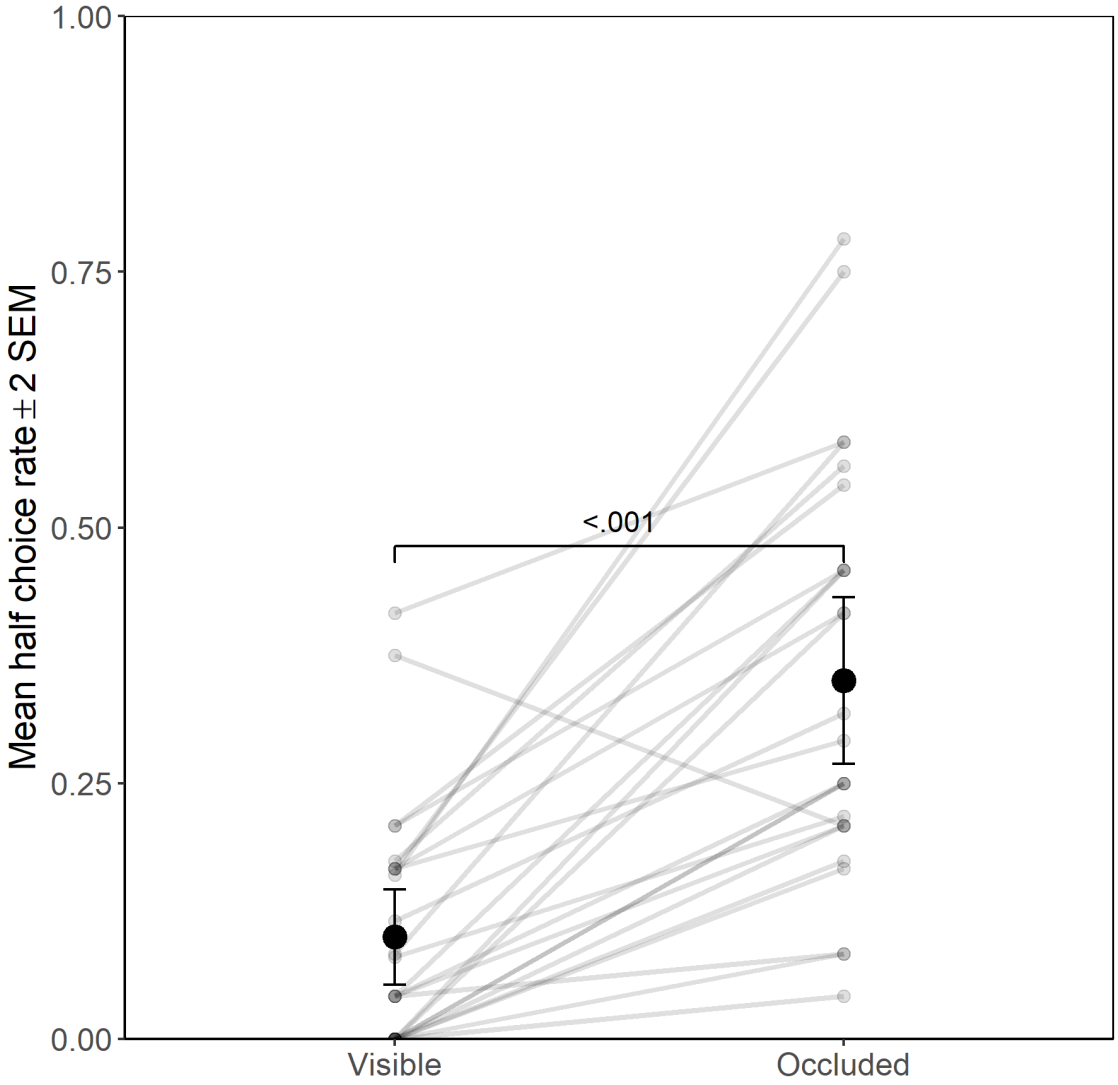


Figure 2.4.1: Group level rates of taking the half grape in visible and occluded trials of experiment 2. Error bars show 2 standard errors around the mean, lines and small points show individual level means.

Table 2.4.2: Individual level differences in half-choice rate by condition in Experiment 2. Fisher's exact test with Holm-Bonferroni correction for multiple comparisons.

Species	ID	Visible	Occluded	Fisher Test (p)	Fisher Test (p.adj)
Bonobo	Cheka	0	0.208	0.025	**0.062**
	Daitou	0	0.417	0.019	0.052
	Likemba	0.167	0.417	0.055	0.099
	Lina	0.417	0.583	0.342	0.389
	Lola	0.083	0.583	0.014	**0.049**
	Lopori	0	0.458	<.001	**0.001**
	Lucuma	0	0.25	0.109	0.151
	Malaika	0	0.25	0.011	0.046
	Ndeko	0	0.042	0.5	0.521
	Rubani	0.16	0.783	<.001	**<.001**
Chimpanzee	Coco	0.174	0.56	0.006	**0.031**
	Holly	0.167	0.292	0.247	0.294
	Josie	0	0.167	0.055	0.099
	Kibali	0	0.083	0.245	0.294
	Samantha	0.208	0.458	0.062	0.104
	Tuli	0.042	0.25	0.049	0.099
Gorilla	Biddy	0.115	0.318	0.086	0.134
	Lope	0.08	0.217	0.175	0.23
	Oumbi	0	0.174	0.046	0.099
	Ozala	0.042	0.208	0.094	0.138
	Shufai	0.042	0.458	0.001	**0.006**
Orangutan	Basuki	0.208	0.542	0.018	0.052
	Batu	0.375	0.208	0.945	0.945
	Kayan	0.167	0.75	<.001	**0.001**
	Maliku	0.042	0.083	0.5	0.521

We fitted a mixed effects model (GLMM) (package: *Lme4*) with a logit link function to predict the likelihood of taking the half piece. We input the random effect of ID alongside the fixed effects of condition (visible/occluded), age (as a polynomial), block (1/2), and species, and the interactions between condition and each of the other fixed effects. The full model fit the data better than a null model with only the random effect (χ^2 = 148.0 df = 13, p < .001), a model without interactions (χ^2 = 20.4 df = 6, p = .003), and a model without the random effect (χ^2 = 26.3 df = 1, p < .001).

The model detected a significant interaction between species and condition (χ^2 = 8.49 df = 3, p = .0368) (Appendix 1, Table 8.1.1), showing that the effect of condition is not equal across species. Paired contrasts from the model (package: *emmeans*) show that orangutans were not

switching differentially based on condition, while the other species are (Table 2.4.2, Figure 2.4.2). We also find an overall effect of block ($\chi^2 = 5.65$, df = 1, p = .017), but not an interaction with condition ($\chi^2 = 0.20$, df = 1, p = .658), showing a general trend towards taking the half piece less frequently in the second block but the effect of condition being equal across both.

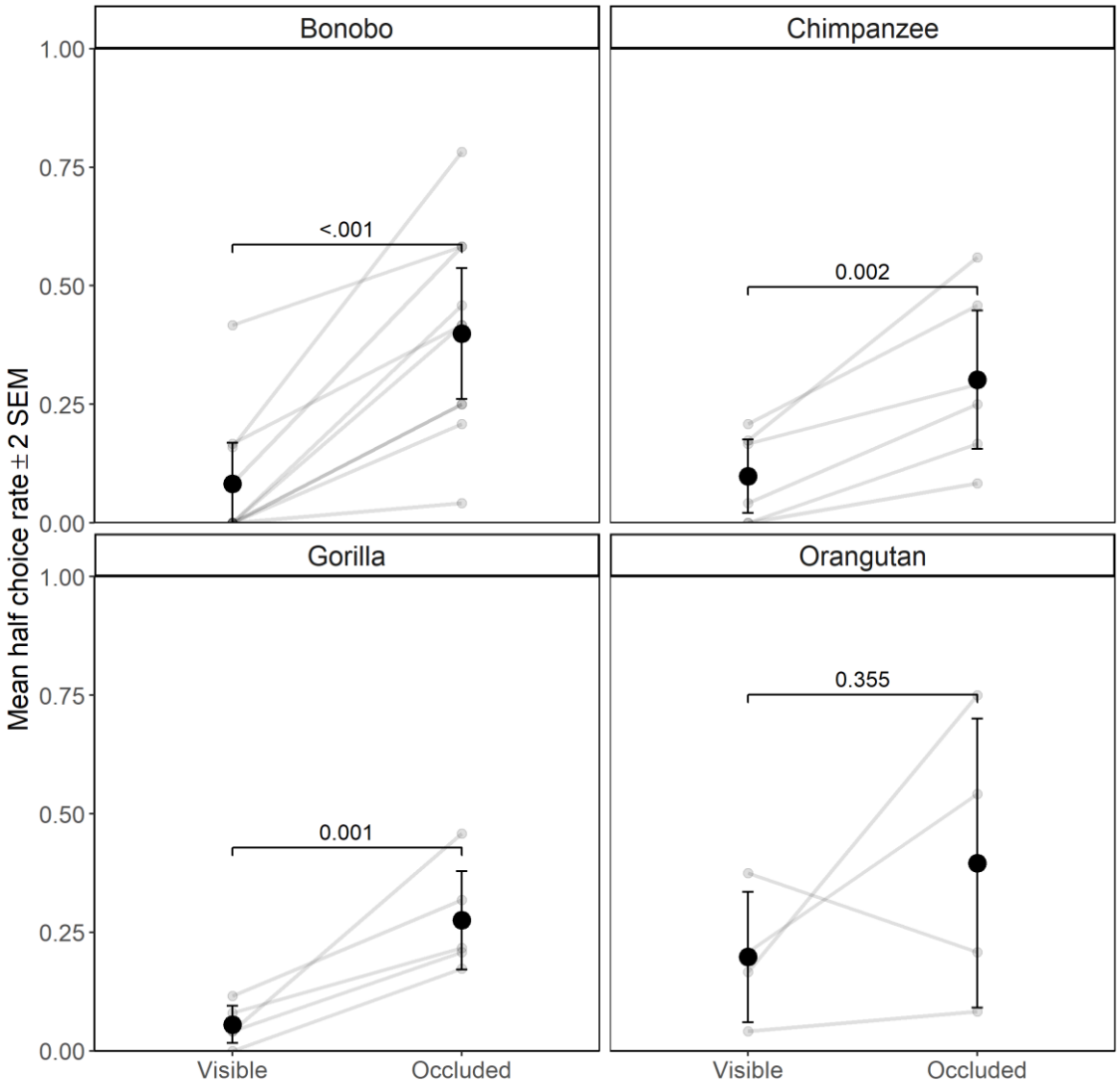


Figure 2.4.2: Mean rate of taking the half piece in experiment 2 by condition and species. Significance tests show pairwise contrasts from the mixed effects model.

Table 2.4.2: Pairwise contrasts between rates of taking the half piece in visible and occluded trials by species.

	β	$CI_{2.5}$	$CI_{97.5}$	p-value
Chimpanzee	-1.601	-0.610	-2.593	0.002
Gorilla	-1.606	-0.615	-2.596	0.001
Orangutan	-0.378	0.423	-1.179	0.355
Bonobo	-1.665	-0.890	-2.440	< . 001

The data also show an interaction between condition and age ($\chi^2 = 12.77$, df = 2 p = .002) (Figure 2.4.3**)**. Upon inspection, we find support for a quadratic relation between half-choice frequency and age in the occluded condition (β = 21.0, CI_{95} (10.4, 31.6), p < .001)[14] but not in the visible condition (β = 8.86, CI_{95} (-3.50, 21.2), p = 0.16) (Table 2.4.3). As this is a measure of confidence, we can interpret these results one of two ways, either that confidence increases during development, peaks during adulthood before decline in old age; or alternatively, the individuals at either end of the age range may be having difficulty in inhibiting selecting the visible option. If this was the case however, then we would expect to see higher rates of taking the half piece in the visible condition, while the effect does trend this way, it does not reach significance. We do, however find support for a linear increase in incorrectly taking the half piece in the the visible condition β = 17.1, CI_{95}(2.58, 31.64) p = .021).

Table 2.4.3: Coefficients from a model to predict taking the half piece in Experiment 2.

	β	$CI_{2.5}$	$CI_{97.5}$	p-value
(Intercept)	-2.221	-3.27	-1.173	<.001
Occluded	2.297	1.152	3.441	<.001
Age	17.116	2.583	31.648	0.021
Age^2	8.859	-3.497	21.215	0.16
Chimpanzee	-0.309	-1.535	0.917	0.622
Gorilla	-0.388	-1.624	0.847	0.538
Orangutan	1.334	0.242	2.426	0.017
Block	-0.295	-0.9	0.311	0.34
Occluded: Age	-19.177	-32.328	-6.026	0.004
Occluded: Age^2	12.1	1.227	22.974	0.029
Occluded: Chimpanzee	-0.064	-1.206	1.078	0.912
Occluded: Gorilla	-0.06	-1.21	1.091	0.919
Occluded: Orangutan	-1.287	-2.254	-0.32	0.009
Occluded: block	-0.163	-0.884	0.558	0.658

Notably, one individual, Coco a 56-year-old chimpanzee, has an age which places her more than 2 standard deviations above the mean of the group, so is technically an outlier. However, this sample is particularly left skewed when compared to other populations tested (e.g. Hopkins et al., 2021) and she is a particularly valuable example for learning about how advancing age impacts great ape cognition, so we have elected to keep her in the analysis. Nevertheless, if we were to remove her and run the model refinement procedure again (Appendix 1, Table 8.1.3),

[14] This coefficient differs from that found in Table 2.4.3, it is instead calculated by changing the reference level to occluded before outputting the coefficients. Coefficients from that model can be found in Appendix 1 Table 8.1.2.

we still find the quadratic relation between age and confidence in the occluded condition (β = 19.2, CI_{95}(8.58, 29.8) p < .001), and a linear relation in the visible condition (β = 13.51, CI_{95}(0.089, 26.932) p = .048).

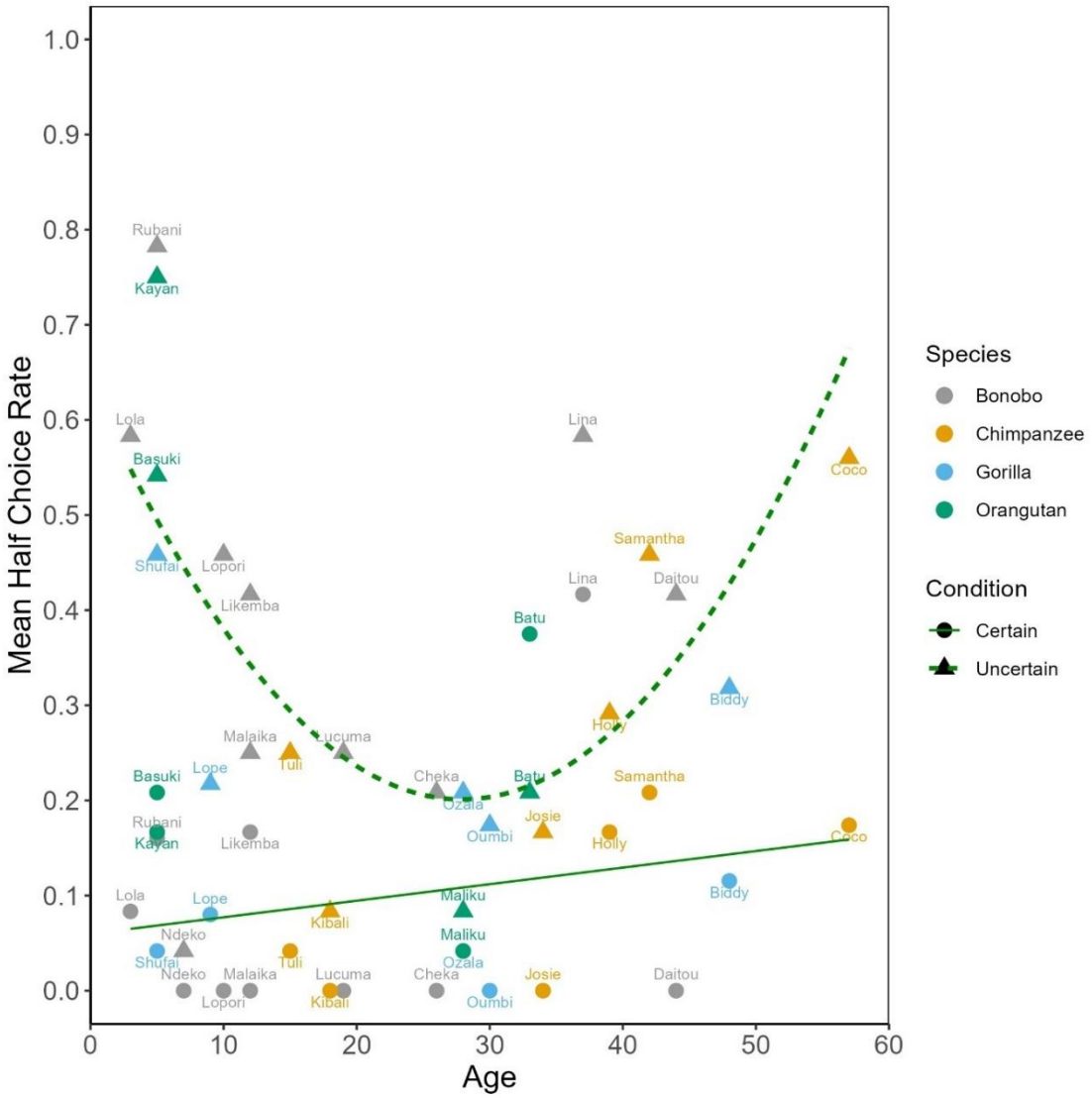


Figure 2.4.3: The relationship between age and taking the half piece in experiment 2 by trial type

A U-shaped relationship between inhibitory control and age is considered to be a healthy developmental trajectory in humans (Diamond, 2013), within the primate literature, there is mixed support for this relationship. In reversal tasks, which train a pre-potent response during training phase but require the subject to abandon it during test trials, great apes show a clear U-shaped relation between age and perseverance with the original strategy (Lacreuse et al., 2018; Marín Manrique & Call, 2015). However, other studies have failed to find age-related effects on inhibitory control in Barbary macaques (*Macaca sylvanus*) (Rathke & Fischer, 2020) or orangutans (Damerius et al., 2017) via an inhibitory reaching paradigm. Notably, inhibitory control of reaching emerges in children and monkeys within the first year (Diamond, 1990), so this may have been too early to have been detected by these studies. However, these differing results may equally be a consequence of task factors as, when given to primates as part of a battery, tasks designed to test inhibition often do not correlate with one another (Völter et al., 2018, 2022).

Figure 2.4.4 shows an overlay of the data from Experiment 1 onto Figure 2.4.3. Notably, while the trend in visible trials is consistent, in occluded trials it is reversed in the Edinburgh chimpanzees, with the middle-aged individuals most likely to take the half grape. This suggests that we should treat that relationship with more caution than that of the visible trials, which was consistent between the two groups.

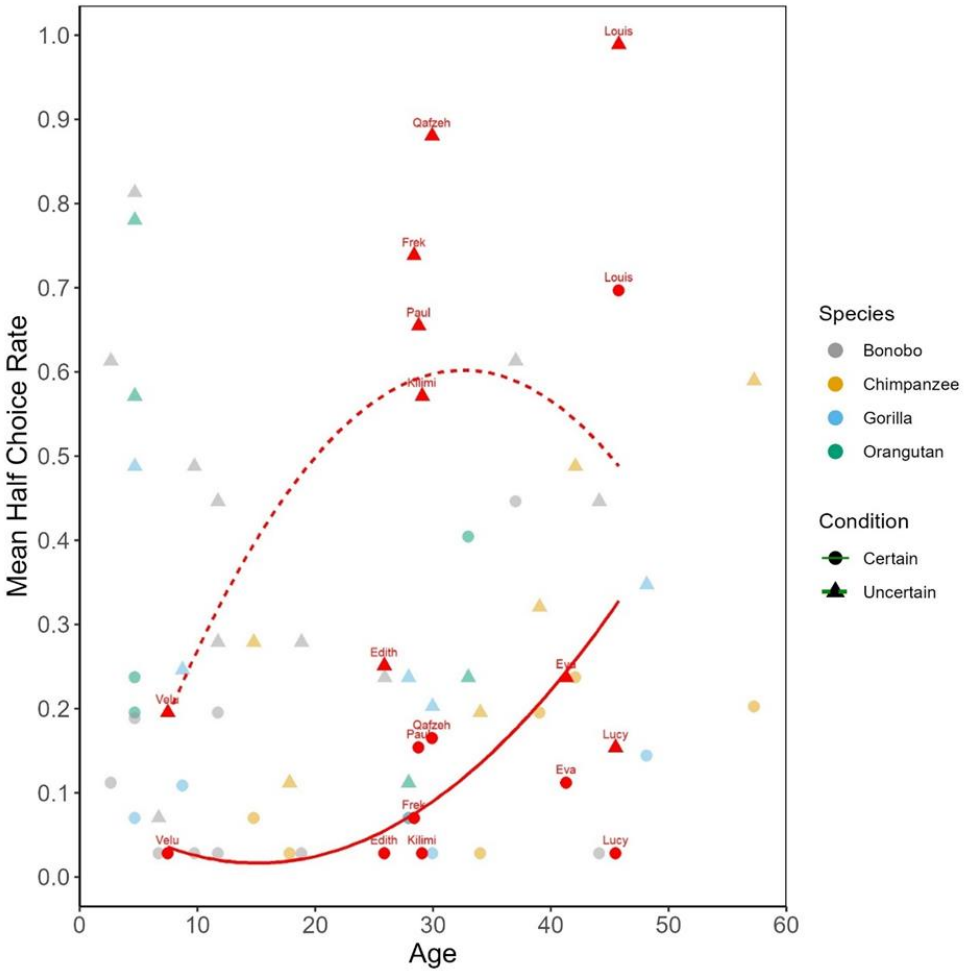


Figure 2.4.4 The relationship between age and half-grape choice for the Edinburgh chimpanzees, superimposed over Figure 2.4.3.

While we have provided strong counter evidence to the sequential guessing hypothesis, Leahy and Carey (2020) also describe how, when faced with a task in which errors are costly, a minimal agent can act as though she is monitoring uncertainty without being aware of her uncertainty. They do so by learning to recognise perceptual features of a presentation which have been previously associated with a decreased frequency of success. To take this more conservative reading of the minimal model, it is possible that subjects recognise the presence of the occluder, associating it with previous uncertainty-based tasks which they have engaged with, and deciding before the point of decision that they would take the half piece without making a simulation. Like our case against the argument made by Redshaw and Suddendorf (2020), testing naïve individuals provides a strong counter to this suggestion, however, it could be argued that these associations had been made in the short length of time which the subjects had engaged with testing, therefore the critique may still stand.

Secondly, the agent may doubt the entire representation as compared to one that is obtained visually, as is the case here, leading to the decrease in confidence, and the decision to opt-out. We will test both of these explanations in Experiment 3, equating the presentations of the conditions to ensure that subjects respond based on the presence of conflicting representations, rather than the strength of the overall representation.

2.5 Experiment 3.

Experiments 1 and 2 tested for a difference in confidence between a simulated representation and one obtained visually. Experiment 3 will use an occluder in both certain and uncertain trials, removing it as a possible clue to uncertainty and equating the representational strength of the uncertain and certain conditions.

Methods

Participants.

Experiment 3 involved the BRU participants from Experiment 1, 10 individuals participated, including two subjects who had not participated in Experiment 1, Liberius, a 24-year-old male who had previously failed the pre-test, and Masindi, a 3-year-old female who had been housed separately during the initial testing period (Table 2.3.1). One individual who had participated previously, Louis, failed to pass the pre-test and was dropped from this experiment.

Apparatus.

The apparatus was the same as used in experiments 1 and 2.

Procedure.

The conditions in experiment 3 were *two-cup*, a replication of the occluded condition of the previous two experiments, and *one-cup*, where baiting took place behind an occluder but there was only one cup on the table. The one-cup condition replaced the visible condition of previous experiments, differing in that although the grape was always located under that cup at the end of the baiting so producing the same level of certainty, the subject did not see it placed there. Subjects received 2 sessions of 12 trials, each containing 6 one-cup and 6 two-cup trials. Which cup was baited in two-cup trials was counterbalanced between trials, as was the position of the single cup (left/right) in one-cup trials. Data were collected approximately 18 months after experiment 1.

Data Coding and Analysis

The experimenter live coded first and second choices, inter-coder reliability based on the first choice, removed cup contents and second choice for 15% of trials was excellent (kappa = .954, n = 32).

Results and Discussion.

Figure 2.5.1 presents the percentage of trials in which subjects took the half grape in each condition, showing a significant difference by condition (t-test, $t_8 = -3.34$, p = .009), replicating the conclusions produced by Experiment 1 but demonstrating that chimpanzees differentiate between two generated representations based on the level of certainty which they produce. This would not be possible under even the more conservative reading of the minimal model. Furthermore, there was no difference in rates of taking the half grape between sessions for either condition (one-cup, $t_9 = -0.194$, p = .851; two-cup, $t_9 = 1.59$. p = .146), thus within the experiment subjects are not learning to associate the two-cup variant with a lower probability of success. Individual rates of taking the half grape by condition can be found in Table 2.5.1.

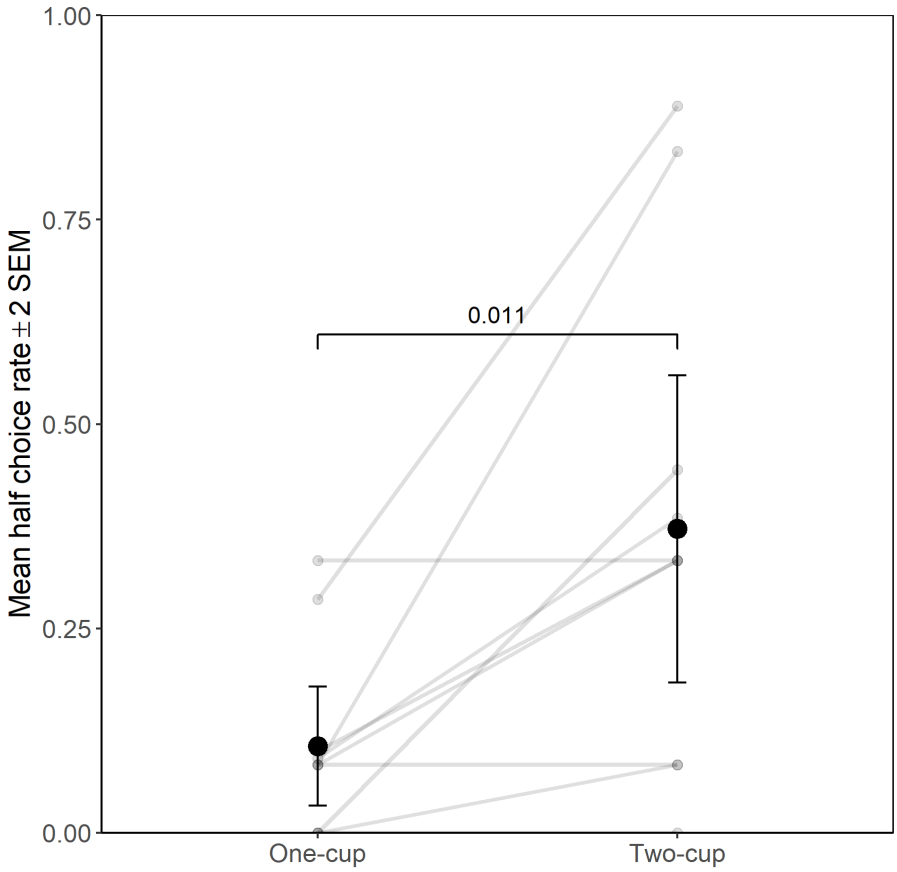


Figure 2.5.1: Group rates of taking the half grape in trials of Experiment 3. Lines and small points show individual mean rates. Significance test shows the resultant p-value from a paired t-test.

Table 2.5.1: Individual half-choice rates in Experiment 3.

ID	Visible	Occluded	Fisher Test (p)	Fisher Test (p.adj)
Edith	0.1	0.333	0.249	1
Eva	0.083	0.083	0.761	1
Frek	0	0.444	0.033	0.26
Kilimi	0.091	0.385	0.118	0.826
Liberius	0.286	0.889	0.024	0.22
Lucy	0	0.083	0.5	1
Masindi	0.083	0.833	0	**0.003**
Paul	0.333	0.333	0.667	1
Qafzeh	0.083	0.333	0.158	0.95
Velu	0	0	1	1

When comparing those individuals who completed both Experiment 1 and Experiment 3 (Figure 2.5.2), there was no difference in the rate of taking the half grape between the visible and the one-cup conditions (t-test, $t_7 = -1.218$, $p = .252$), which suggests that chimpanzees are treating them equally.

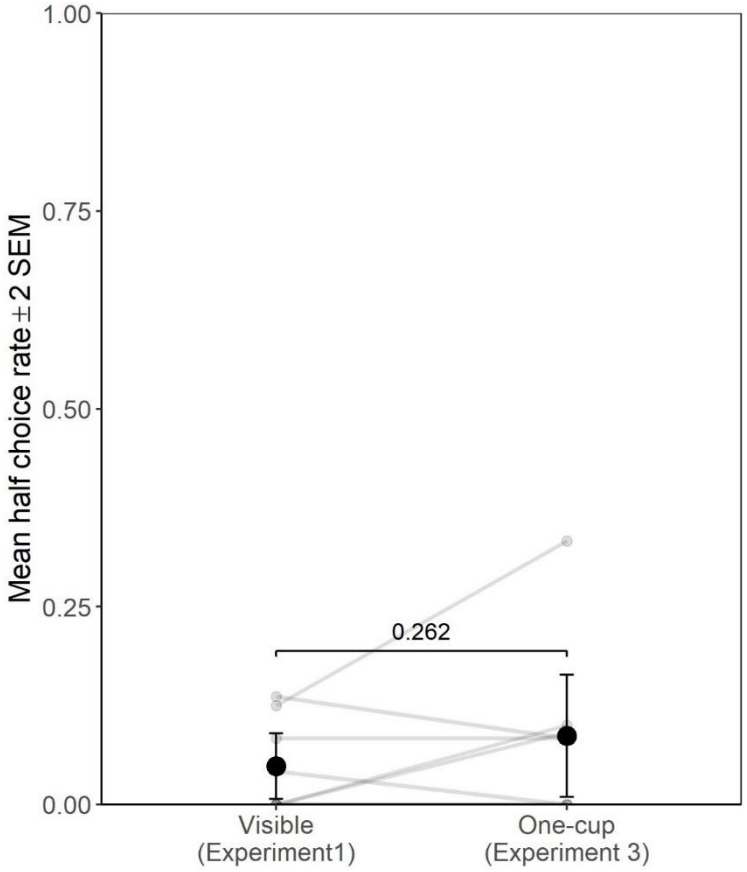


Figure 2.5.2: Difference in half-choice rates in certain trials of Experiment 1 and Experiment 3.

Engelmann et al. (2023) frame the location-based argument to explain the performance of chimpanzees in the 3- and 4-cup tasks, proposing that subjects mark a broad location for each item, covering the range of cups it could be under, when a cup is revealed the subject can shrink or eliminate a location, but if both remain at the point of decision they will pick indiscriminately between the two locations. This leads them to choose equally between the single cup, certainly containing a grape, and the pair of cups, in which each only has a 50% chance of containing a grape. In our task, the one-cup condition is theoretically equivalent to the certain cup in the 3- and 4-cup paradigms, while the two-cup condition is equivalent to the uncertain pair. If these two are valued equally, then we would see them chosen at the same rate against the half-piece, which we do not, instead we see the subjects choosing in line with expected value.

2.6 General Discussion

Across three experiments we have provided evidence to counter the conclusion that great apes only consider one possibility when making their decisions. Experiment 1 showed that chimpanzees do not treat their guesses as though they are certainties, which shows that they are not simply engaging with a process of sequential guessing. Experiment 2 extended this paradigm to the other great ape species, finding that this capacity is shared across the hominid lineage, but that it varies systematically with age. Finally, we tested a more conservative version of the minimal model of possibility by equating the representational demands of the certain and uncertain condition, demonstrating that the effect still holds under this more exacting standard.

While the extended time period between the two studies and the different individuals participating mean that comparisons should be treated with caution, the mean rate of taking the half-grape in one-cup trials of Experiment 3 was comparable to visible trials of Experiments 1 and 2. This is important as it would suggest that participants treat the two conditions equally, while we cannot definitively measure certainty, treating the inferred location as equivalent to visual evidence is a valuable validation of our initial task.

Crucially, in Experiment 3, when choosing against an alternative with constant value, the half grape, subjects altered their choice rate between conditions based on how many cups were behind the barrier during baiting. This suggests that they are valuing the one-cup and the two-cup conditions differently. This not only refutes the location-based argument as discussed above, but also the ratio of ratios account (Eckert, Call, et al., 2018; Hanus & Call, 2014), demonstrating that apes are actually able to distinguish between $p = 1$ and $p = 0.5$. Possibly, it is only when directly choosing between two uncertain outcomes that this is the case, whereas in our paradigm, because the decisions take place sequentially, the status of the first choice, be it

certain or uncertain, is added to the subject's model of the world, and then the subject makes the choice between it and the half grape. We could hypothesize that, if we repeated the 3-cup paradigm but the choice was offered sequentially as we have done here, apes would continue choosing in line with expected value and we would see a higher proportion of choices towards the certain grape.

Throughout the three experiments we see approximately 10% of choices directed towards the half-piece on certain trials, thus echoing Hanus and Call's (2014) conclusion that apes fail to recognise the special status of a certain outcome. This partially echoes the conclusion of Leahy and Carey (2020), without the word to describe certainty, non-linguistic groups lack the understanding of the concept. This is reminiscent of the divide between inferential and deductive reasoning, while inferential reasoning is probability based, deductive reasoning is a wholly language-based concept. I support the conclusion that apes live their lives probabilistically, as do children, and will expand upon this point in future chapters.

3. Disjunctive reasoning: ruling out the impossible.

The data presented in experiments 1 and 3 of this chapter have been published as part of the following paper:

Jones B, Call J. Chimpanzees (Pan troglodytes) recognize that their guesses could be wrong and can pass a two-cup disjunctive syllogism task. Biol Lett. 2024 Jun;20(6):20240051. doi: 10.1098/rsbl.2024.0051.

3.1 Abstract

Chimpanzees and young children consistently struggle with tasks which test the disjunctive syllogism, A *or* B *not* A *therefore* B, which has led some authors to argue that language is a necessary pre-requisite to solving these tasks. We modified our post-decision wagering paradigm to test the disjunctive syllogism by giving subjects information about the unchosen cup before they chose between their cup and the fractional reward, finding that chimpanzees were able to flexibly adjust their choice behaviour accordingly. In experiment 2 we extended this finding to a naïve cohort, showing that the effect was both consistent across species and not a consequence of the first group's previous experience. In experiment 3, we added additional controls for non-cognitive strategies and found the same effects. These data suggest that language is not a pre-requisite to solving the disjunctive syllogism and provide a valuable contribution to the debate on logical reasoning in non-human animals.

3.2 Introduction

Chapter 2 has demonstrated that great apes possess more than simply a minimal model of possibility. However, a mature model also requires that a thinker represents these possibilities as mutually exclusive. Importantly, the reasoner must represent the likelihoods of each possibility as dependent on one another. For example, in a search task each failure to find an item increases the likelihood of it being found in a yet to be searched location. The most minimal example of which is the disjunctive syllogism: A *or* B, *not* A *therefore* B or equally A *therefore not* B. This can be more simply referred to as inference by exclusion.

As discussed in the introduction, inference by exclusion has a long history in the comparative literature, with the general consensus that primates are able to pass a 2-cup task. However,

Paukner et al. (2006) propose a process of simply avoiding unbaited containers, to explain why, in a metacognition paradigm, capuchins fail to choose differentially between 2 unbaited tubes and a bent tube, which could be baited but not confirmed by visual inspection. This later became known as the *avoid the empty cup hypothesis*. Paukner followed up this with a second experiment with capuchins showing that they did have a strong tendency to avoid an empty cup (Paukner et al., 2009), and Schmidt and Fischer (2009) note that this account can explain the performance of some, but not all, baboons in an inference by exclusion task.

Call (2022) tested this hypothesis explicitly in great apes using a 3-cup 1-item paradigm. He baited a pair of cups behind a barrier, while a third cup rested on the table untouched, then showed the subject that one of the possibly baited cups was empty and gave the subject a choice between the 3. In the test condition subjects selected the cup that had been behind the barrier above chance levels, while in a control condition where the experimenter did not reveal the contents of the empty cup, they were at chance. In a second experiment in which the experimenter instead removed the baited cup, the apes showed no preference for the cup behind the barrier during baiting, suggesting that subjects treated it equally to a cup which had zero chance of having been baited. So the evidence would suggest that, for apes at least, subjects can solve a 2-cup task inferentially.

However, this capacity is seemingly not extensible to the 4-cup task, in which subjects are presented with 2 pairs of cups, each baited with one item. When shown that one cup in one of the pairs is empty, chimpanzees only choose the other cup in the pair on 50% of trials, but when they see that the food item is taken away, they switch on ~ 85% of trials (Engelmann, Haux, et al., 2023). This would suggest that chimpanzees are treating the *or* in the disjunctive syllogism as *exclusive*, that it cannot be both under A and under B, but not *inclusive*, that it must be under A or under B. The authors go on to formulate their location-based argument (as described in the introduction and previous chapter), a quasi-logical explanation for the chimpanzees being able to rule out where a grape *is not* and avoid it, but not making expectations about where the grape *is*. This is the opposite of the relationship in children, who treat the or as *inclusive* between the ages of 2.5- and 5-years-old, and only recognise or's *exclusive* meaning after the age of 5 (Gautam et al., 2021b).

The current work

This chapter aims to test whether, under the same 2-cup post-decision wagering paradigm as chapter 1, great apes are able to reason via both variants of the disjunctive syllogism. Experiment 1 tests this in the Edinburgh chimpanzees, experiment 2 repeats the procedure with

the Twycross apes and experiment 3 tests the Edinburgh group under a revised paradigm which rules out non-cognitive strategies for solving the task.

3.3 Experiment 1

Methods

Participants.

We tested 8 chimpanzees (4 female) aged between 7 and 46 (mean age = 31.7 years), detailed demographic data can be found in Table 3.3.1. Subjects were housed at the Budongo Research Unit (BRU), which operates within Edinburgh Zoo. The subjects live in a natural group, enclosures allow access to both indoor and outdoor space with vegetation. The chimpanzees receive regular feedings throughout the day which are comprised of a wide variety of fruits and vegetables, the group additionally receives further enrichment. Individuals are experienced in non-invasive cognitive testing and similar search paradigms. Testing is voluntary, non-contact and takes place in a communal area accessible to all group-members, at no point were subjects separated from their group.

Table 3.3.1: Demographic details of participants in experiments 1 and 3.

			Age	
ID	Sex	Rearing	Exp 1	Exp 3
Edith	Female	Parent	25	26
Eva	Female	Parent	41	42
Frek	Male	Parent	28	29
Kilimi	Female	Parent	29	30
Louis	Male	Wild-caught	45	-
Lucy	Female	Parent	45	46
Paul	Male	Parent	-	29
Qafzeh	Male	Parent	30	31
Sophie	Female	Parent	-	41
Velu	Male	Parent	7	8

Apparatus

The apparatus was the same as that used in experiments 1 and 3 of chapter 2. A sliding table (630mm x 300mm) was attached on the outside of the chimpanzees' enclosure, 3 holes at the base of the plexiglass panel allowed the subject to indicate their choice by placing one of their fingers into the hole. The same two cups (Ø = 86mm, height = 89mm) were used as baiting locations and the same occluder (height = 250mm, width = 450mm) was used to conceal the baiting.

Procedure.

Test trials.

Experiment 1 followed on directly from Experiment 1 in the previous chapter with no additional training trials. The experimenter sat opposite the subject with a sliding table placed between them, at the back of the table were two identical cups. To start the trial the experimenter lifted the cups to show that they were empty and placed them at the front of the table, which was still outside of reach of the subject, and then placed an occluder in front of the cups. The experimenter held a whole grape above the occluder before baiting one of the cups with it and showing their open hands to the subject. The experimenter visited both cups during the baiting, both the order visited, and the hiding location of the grape were counterbalanced between trials. The experimenter then removed the occluder and slid the table to its forwards position to allow the subject to indicate their choice. Once the subject had indicated their choice, the experimenter moved the table to its backwards position and while touching both cups, they lifted the unchosen cup and moved it to the back of the table. If the unchosen cup had been baited, they removed the grape and placed it into a bucket on the floor while the subject watched, but if it had been empty, the experimenter looked at the spot originally occupied by the cup for 3 seconds. These conditions are referred as to *reveal baited* and *reveal empty*, respectively. In both conditions the experimenter then placed a half-grape on the original position of the nonchosen cup before offering the subject a choice again by sliding the table to its forward position. Figure 3.3.1 shows a diagram of the procedure of experiments 1-3.

Subjects received 24 trials split into two equal blocks, assuming that they are unable to detect the location of the whole grape they should receive an equal number of each condition.

Data scoring and analysis

The experimenter live coded the first choice, the contents of the revealed cup and the subject's second choice. A second experimenter recoded 15% of trials from video recordings, inter-observer reliability was excellent (kappa = .959, n = 29). Analysis was based on the rate of taking the half-grape in reveal empty and reveal baited trials. All analysis was completed in R (version 2021.09.1).

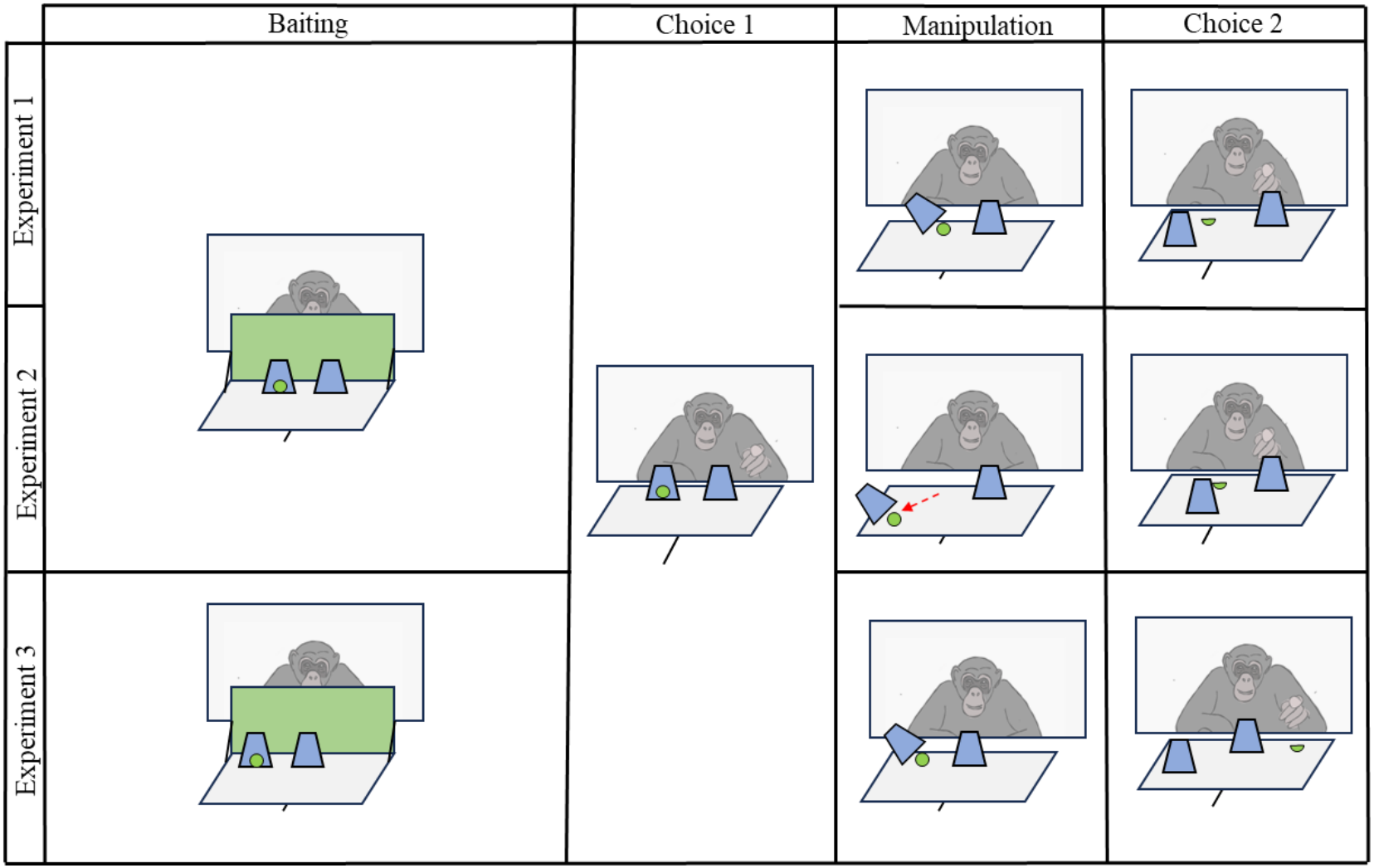


Figure 3.3.1: Procedure for test trials of Experiments 1-3. Adapted from Jones and Call (2024). Visible trials of experiment 2 were identical but without the occluder.

Results and Discussion

From 192 trials, 95 were *reveal baited* and 97 *reveal empty.* Figure 3.3.2 presents the percent of trials in which subjects took the half-grape as a function of the contents of the revealed cup, subjects chose the half grape more frequently on *reveal baited* than *reveal empty* trials, but not significantly so (t-test, $t_7 = -1.61$, $p = .152$). When compared to occluded trials of Chapter 1 Experiment 1, which acts as a *no information* comparison, we find that subjects failed to adjust their half-choice rates adaptively in either condition (reveal empty: paired t-test (one-tailed), $t_7 = 1.76$, $p = .061$; reveal baited: paired t-test (one-tailed), $t_7 = -.209$, $p = .420$).

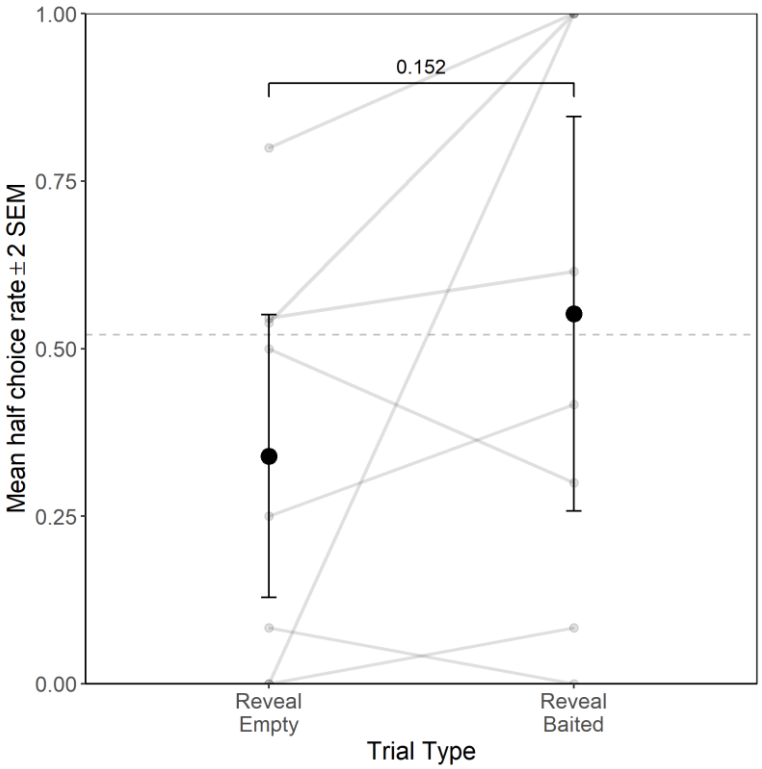


Figure 3.3.2: Group level and individual level rates of taking the half grape by trial type in experiment 2. The hashed line shows the group level rate of taking the half grape in a no-information condition (Chapter 2, Experiment 1, occluded trials). Annotations show the resultant p-value from a paired t-test for differences by trial type.

To test for learning and main effects, we fitted a GLMM model (package: *lme4*) to predict the binary outcome of choosing the half grape (Table 3.3.2), using condition, session and the condition-session interaction as fixed effects, and individual ID as a random effect. The random effect of individual improved the fit over a GLMM including only the fixed effect structure ($\chi^2 = 59.63$, $df = 1$, $p < .001$) and the GLMM that included the fixed effects was an improvement over

the null model containing only the random effect (χ^2 = 27.5, df = 3, p < .001)[15]. The interaction term indicated diminished effect of condition in the second session, specifically through a reversion to chance in the reveal empty condition (Figure 3.3.3). When we reanalyse only the first session, with an adjusted alpha of .025 to account for multiple comparisons, we see a significant difference between the conditions in the first session (t-test, t = -2.99, df = 7, p = .020).

Table 3.3.2: Coefficients from the mixed effects model to predict the likelihood of taking the half grape in Experiment 1.

Term	β	95% CI		p-value
(Intercept)	0.403	-1.741	2.548	0.712
Reveal Empty	-4.722	-7.343	-2.101	<.001
Session	-0.084	-1.124	0.955	0.874
Reveal Empty: Session	2.188	0.611	3.764	0.007

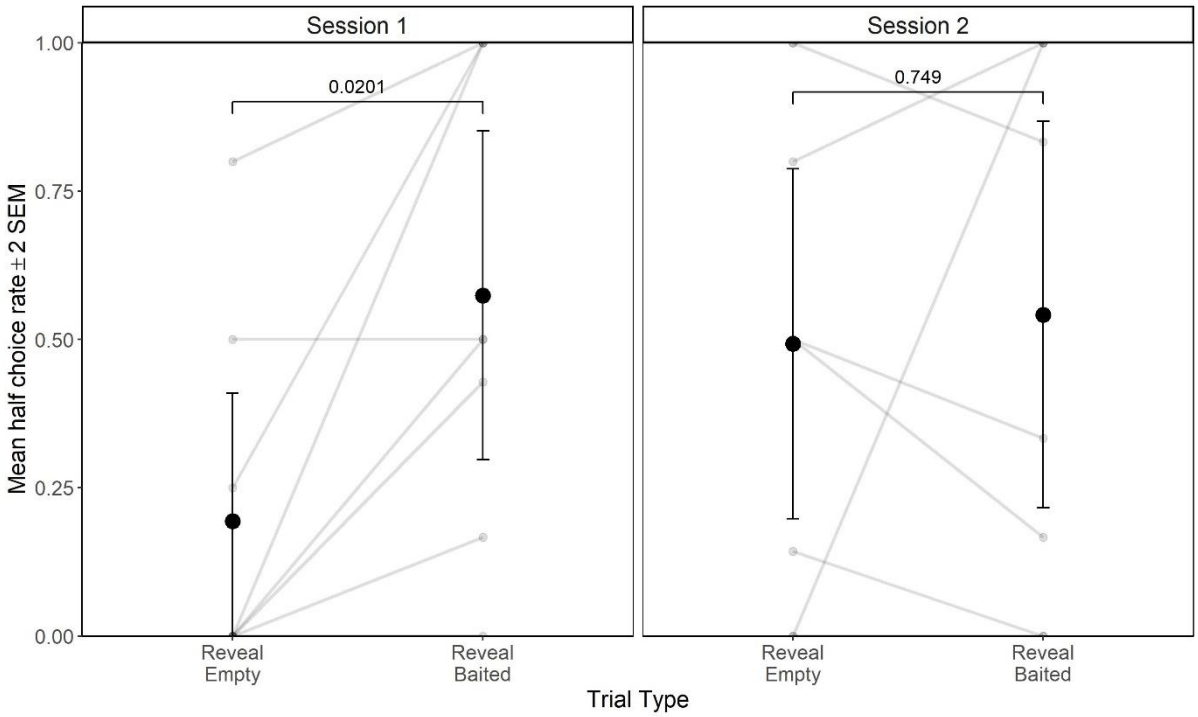


Figure 3.3.3: Group- and individual rates of taking the half grape by trial type and session. Annotations show the resultant p-value (unadjusted) from a paired t-test for differences by trial type.

[15] The inclusion of individual base rate from the no-information condition (Chapter 2 Experiment 1), did not improve the fit of the model (χ^2 = 4.44, df = 2, p = .108)

Table 3.3.3 shows individual rates of taking the half grape by condition. Fisher's exact tests (one-tailed) revealed a significant relation between the contents of the revealed cup and choosing the half grape for two individuals in Experiment 1, Frek ($p = .013$) and Velu ($p < .001$), who also both adapted their rate from Experiment 1 in the predicted direction for each condition.

Table 3.3.3: Individual half choice rates for experiment 1. Fisher's exact test, with Holm-Bonferroni correction for multiple comparisons.

ID	Reveal empty	Reveal baited	Fisher Test (p)	Fisher Test (p.adj)
Edith	0.545	0.615	0.527	0.702
Eva	0	0.083	0.5	0.702
Frek	0.538	1	0.013	0.051
Kilimi	0.25	0.417	0.333	0.667
Louis	0.5	0.3	0.92	1
Lucy	0.083	0	1	1
Qafzeh	0.8	1	0.163	0.435
Velu	0	1	0	<.001

Notably, Velu answered correctly on all 24 trials of both conditions. The two conditions test the inclusive-, A *or* B, *not* A *therefore* B, and the exclusive disjunction A *or* B, A *therefore* not B. Children can pass the inclusive disjunction at the age of 2½ but cannot pass the exclusive disjunction until the age of 5 (Gautam et al., 2021b). The success of one individual negates the suggestion that language is a necessary pre-requisite for solving the disjunctive syllogism, however, it may also reflect that that individual may have used a non-inferential strategy. Individual differences are characteristic of comparable primate literature (Call, 2022; Engelmann et al., 2021; Engelmann, Haux, et al., 2023; Ferrigno et al., 2021) and inference ability has been suggested to be one of the scales upon which primate cognition varies (Herrmann & Call, 2012).

Finding comparable results in multiple individuals will bolster the idea that this capacity is not language dependent. In Experiment 2 we extended the paradigm to the other great ape species and to a larger, more diverse, cohort. Crucially, unlike the Edinburgh chimpanzees these individuals are not research experienced and have never participated in any inference tasks, thus they cannot bring associative rules or knowledge from outside of their standard lived experience. This means that, as close as possible, they reflect an unadulterated sample, and their performance could be considered species typical.

3.4 Experiment 2.

Methods

Participants.

We tested 21 apes housed at Twycross Zoo, 6 bonobos, 6 chimpanzees, 5 gorillas and 4 orangutans. (13 female, mean age = 22.3). Subjects are housed in species-typical groups and testing takes place voluntarily in a communal area, further details of housing along with demographic details of the participants can be found in Experiment 2 of Chapter 2. This experiment directly followed from that one. Although there were no additional pre-test trials before this experiment, 4 bonobos (*Diatou, Lina, Lola, and Lucuma)* who previously participated, did not take part in this experiment.

Apparatus.

The table, cups and occluder were the same as Experiment 1. In place of grapes the food items were pieces of raw sweet potato, the large piece was twice the volume of the fractional piece, $2cm^3$ and $1cm^3$ respectively. For one individual, Batu, 2 fractional pieces were used in place of the large reward due to him not reliably selecting the large piece during the initial training phase.

Procedure.

The procedure for test trials was the same as Experiment 1, with the exception that we included a visibly baited trial every third trial, to determine whether apes treated a cup they had inferred was baited as equivalent to one they had seen baited. Older children's equivalent treatment of inferences and visual evidence is used by Mody and Carey (2016) to evidence their conclusion that children are reasoning 'logically' - which implies deductive reasoning. Finding the same in great apes would allow us to draw the same conclusion

Data scoring and analysis

Data was coded as in experiment 1, inter-observer reliability based on 15% of trials was excellent (kappa = .966, n = 111).

Results and Discussion.

In occluded trials subjects guessed correctly on their first guess in 50.5% (±1.54%) of trials, resulting in approximately equal proportions of *reveal empty* and *reveal baited* trials. Figure 3.4.1 presents the percent of trials in which subjects took the half piece as a function of trial type.

Subjects took the half piece on 61.6% of *reveal baited* trials and 23.6% of *reveal empty* trials. This difference was significant (t-test, t_{21}= 5.373 p <. 001). Subjects rationally adapted their half choice rates from the no information condition (Chapter 2, Experiment 2) for both conditions (reveal empty: t-test, t_{20}= 2.26 p =.035; reveal baited, t-test, t_{20}= -5.90, p <. 001), and switch rates in both conditions were different from chance (reveal empty: t-test, t_{20}= -4.44, p < .001; reveal baited, t-test, t_{20}= 2.09, p = .049). Taken together, these results demonstrate that, in a 2-cup paradigm, a naïve cohort adapts their choice behaviour in line with reasoning via both variants of the disjunctive syllogism, which requires treating the OR relation as both *inclusive*, certainly A or B, and *exclusive*, A or B not both.

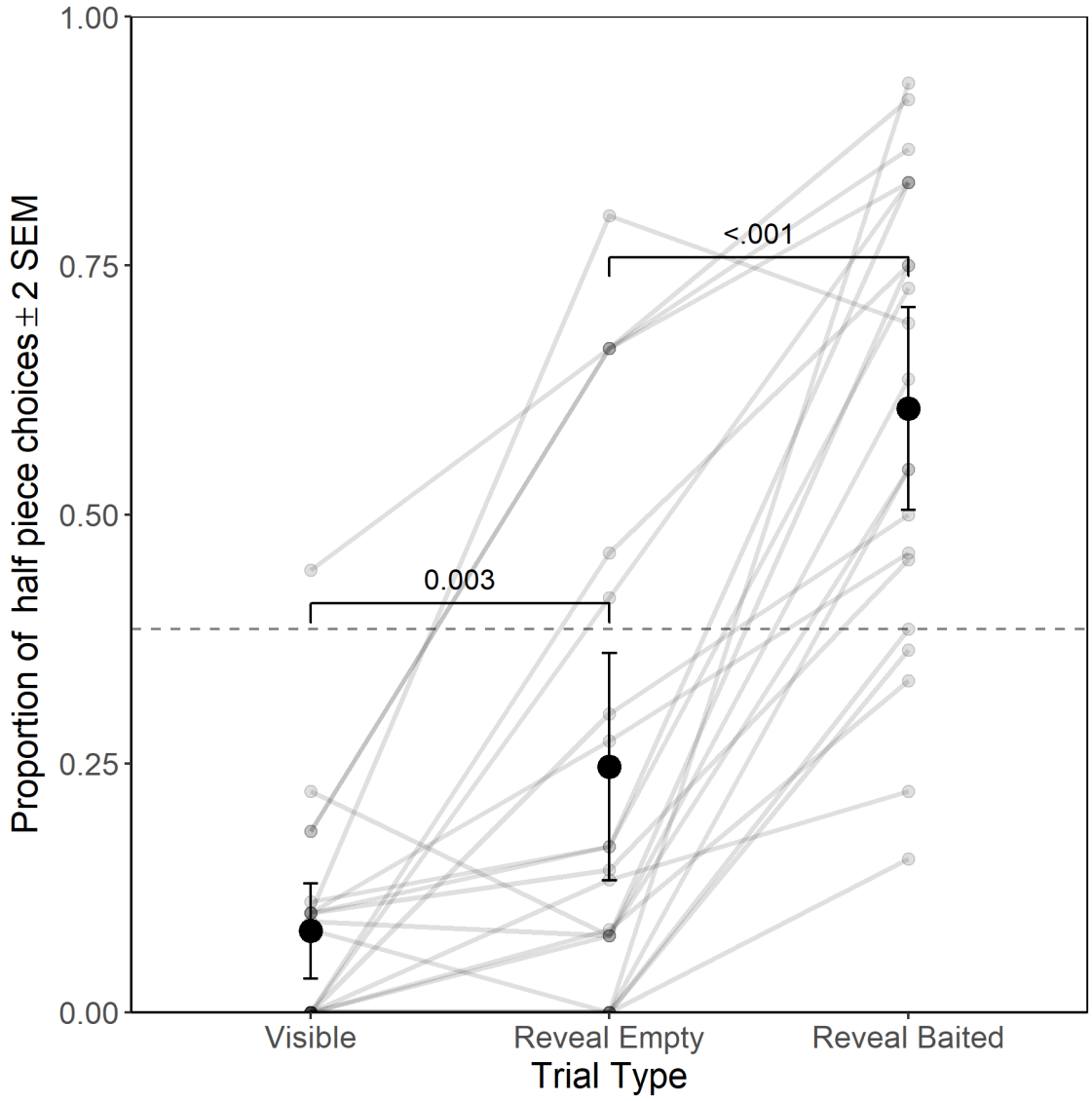


Figure 3.4.1 Proportion of trials in which subjects took the half piece in Experiment 2 as a function of condition, the hashed line shows the group level rate of taking the half piece in occluded trials of Experiment 1 sessions 1 and 2. Analysis of the visible condition only includes those trials where subjects chose correctly on their first choice.

However, when comparing reveal empty trials with visible trials (Figure 3.4.1) apes took the half grape significantly more often in the reveal empty trials (t-test, $t_{21} = -3.35$, $p = .003$)[16], which suggests that they are not treating their inferences as equivalent to visual evidence and that we should not consider their inferences deductive.

At an individual level, six individuals adaptively adjusted their half-piece rates between conditions (Table 3.4.1), and one individual, Kibali, an 18-year-old male, answered correctly on 96% of trials, replicating the performance an exceptional individual in experiment 1. When we consider only the first trial of each condition, 5 of 21 individuals took the half piece on their first reveal empty trial and 10 of 21 did so on their first reveal baited trials. While this difference was not significant ($\chi^2_{42} = 1.66$, $p = .198$), analysing the sessions separately as we did in experiment 1 we see that subjects are adjusting their half-choice rates adaptively from the first session (Figure 3.4.2), which would suggest that this is not a learned association.

Table 3.4.1: Individual rates of taking the half piece in experiment 2

Species	ID	Visible Trials	Reveal Empty	Reveal Baited	Fisher Test (p)	Fisher Test (p.adj)
Bonobo	Cheka	0	0.133	0.222	0.307	0.323
	Likemba	0.091	0.077	0.75	0	0.001
	Lopori	0	0.3	0.5	0.058	0.076
	Malaika	0	0.077	0.636	0	0.001
	Ndeko	0	0.083	0.333	0.034	0.058
	Rubani	0.1	0.8	0.692	0.306	0.323
Chimpanzee	Coco	0.182	0.667	0.867	0.019	0.039
	Holly	0.1	0.143	0.455	0.036	0.058
	Josie	0	0	0.154	0.144	0.178
	Kibali	0.083	0	0.933	0	0
	Samantha	0	0.462	0.75	0.041	0.062
	Tuli	0	0	0.545	0	0.001
Gorilla	Biddy	0.1	0.273	0.462	0.26	0.303
	Lope	0.182	0.667	0.833	0.033	0.058
	Oumbi	0	0	0.385	0.045	0.064
	Ozala	0.1	0.167	0.727	0.001	0.002
	Shufai	0	0.417	0.833	0.001	0.002
Orangutan	Basuki	0.111	0.167	0.833	0.001	0.002
	Batu	0.222	0.077	0.545	0.01	0.025
	Kayan	0.444	0.667	0.917	0.484	0.484
	Maliku	0	0	0.364	0.012	0.028

[16] This analysis only includes trials where subjects correctly chose the baited cup on their first choice.

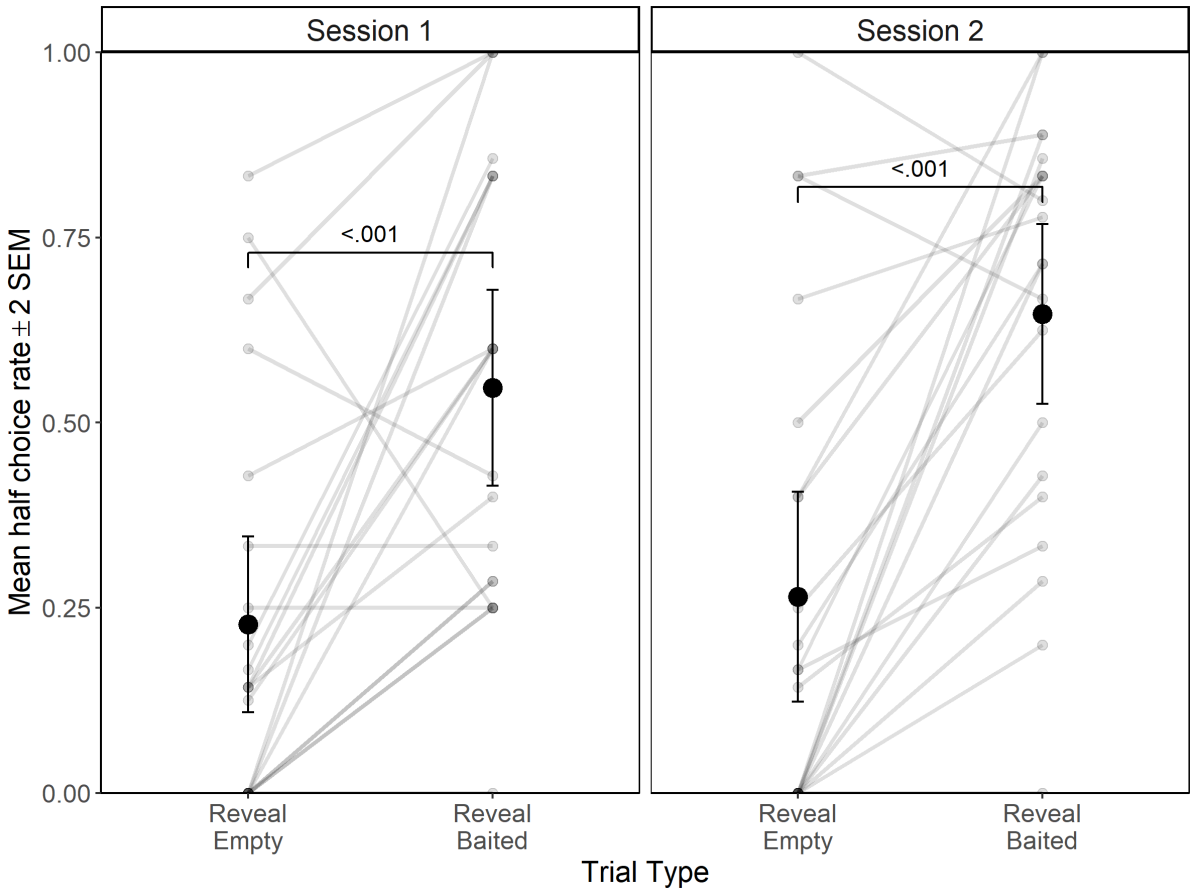


Figure 3.4.2: Half choice rates by condition and session for experiment 2.

We fitted a mixed effects model (GLMM) with a logit link function to predict the likelihood of taking the half piece in Experiment 2. Alongside the random effect of ID, we input the fixed effect of trial type (reveal baited/reveal empty), the subject's *base rate* of taking the half piece in the no-information condition (Chapter 2, Experiment 2), age, and trial number (1-24), along with the interaction between trial type and each of the other fixed effects[17](Table 8.2.1).

We found no support for an interaction between trial type and trial number ($\chi^2 = 0.510$, df = 1 p = .475), meaning that we can rule out this being a learned response. There was a borderline significant interaction between trial type and species, $\chi^2 = 7.807$, df = 3, p = .050), reflecting a larger effect of condition in orangutans ($\beta = -1.90$, $CI_{95}(-3.47, -0.33)$) (Figure 3.4.3).

[17] A model which included the 3-way interaction between trial-type, base rate and age failed to meet convergence criteria.

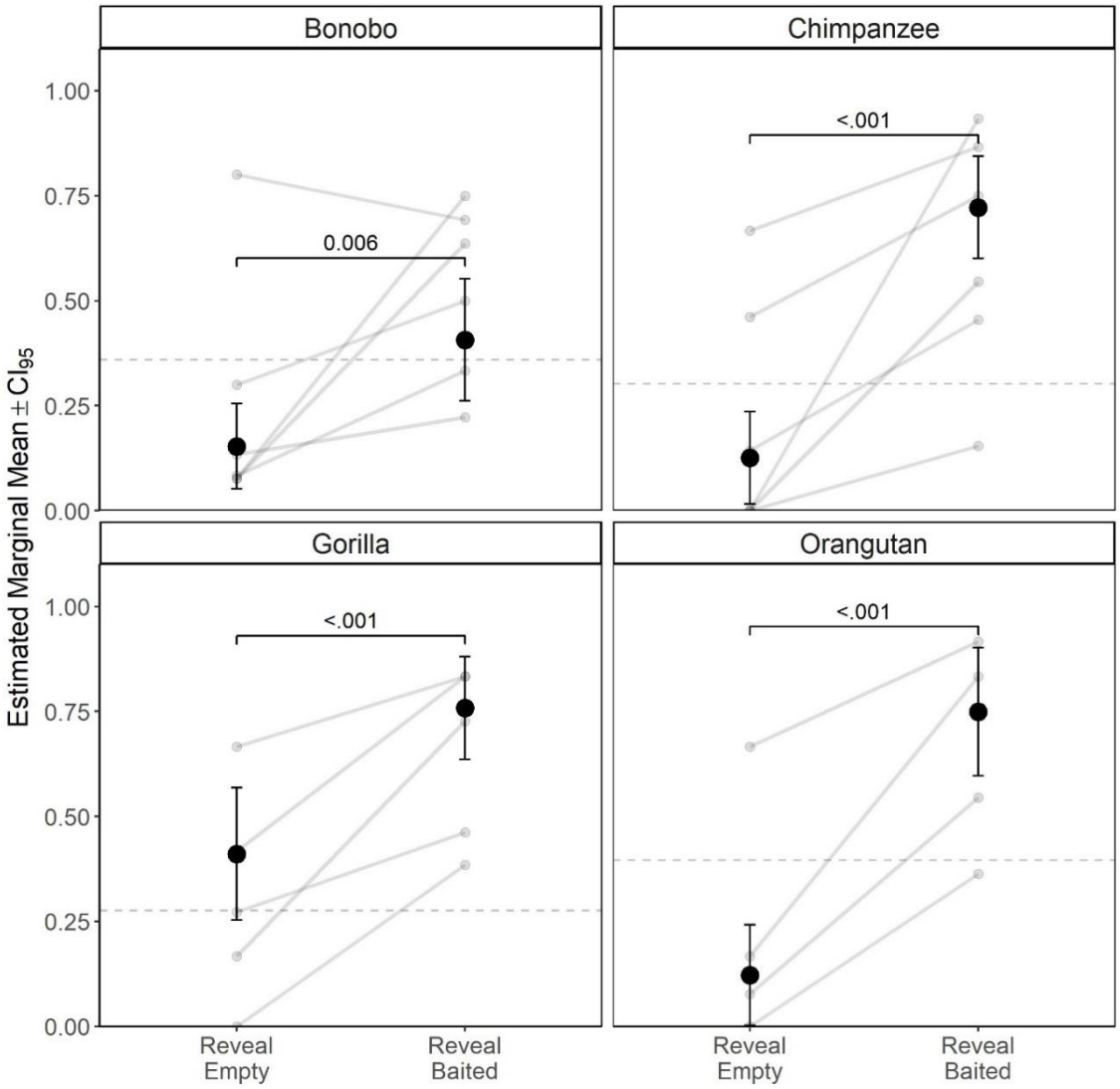


Figure 3.4.3: Estimated marginal means from the model to predict taking the half-piece in Experiment 2. Significance tests show pairwise contrasts from the model. Points and lines show individual mean rates across the 4 sessions.

We do find both a main effect of base rate ($\chi^2 = 13.35$, df = 1, $p < .001$) and an interaction with trial type ($\chi^2 = 6.98$, df = 1, p = .008). The effect of base rate suggests that it is not simply a measure of confidence or risk tolerance which is causing the half-choice in the no information condition. When we examine the interaction between trial type and base rate, we find that the base rate has a more pronounced effect in reveal empty trials ($\beta = 2.53$, $CI_{95}(0.653, 4.41)$, $p < .001$), where the subject has to inhibit taking the half piece, than in reveal baited trials ($\beta = 1.82$, CI_{95} (0.237,3.40)) (Figure 3.4.4). As those individuals with a higher base rate show a smaller effect of condition, this lends support to the second hypothesis developed in Chapter 2, that it is an aspect of executive function causing this base rate change, rather than individual risk propensity or confidence in one's answer.

*Table 3.4.2: Coeffcients,from a model to predict taking the half-piece in experiment 2. (half_choice ~ trial-type*base rate + trial-type*age + trial-type*species + trial-type*trial-number, family = binomial(link = "logit"))*

	β	$CI_{2.5}$	$CI_{97.5}$	p-value
(Intercept)	-0.530	-1.726	0.667	0.386
Reveal empty	-2.650	-4.142	-1.158	<.001
Age	-0.032	-0.066	0.001	0.059
Base rate	1.819	0.238	3.400	0.024
Trial Number	0.025	-0.016	0.065	0.234
Chimpanzee	1.331	0.018	2.645	0.047
Gorilla	1.222	-0.035	2.479	0.057
Orangutan	0.776	-0.524	2.077	0.242
Reveal empty: Age	0.031	-0.004	0.067	0.086
Reveal empty: Base rate	2.529	0.653	4.406	<.001
Reveal empty: Trial Number	-0.023	-0.087	0.041	0.475
Reveal empty: Chimpanzee	-1.323	-2.854	0.208	0.090
Reveal empty: Gorilla	-0.291	-1.597	1.014	0.662
Reveal empty: Orangutan	-1.903	-3.471	-0.334	0.017

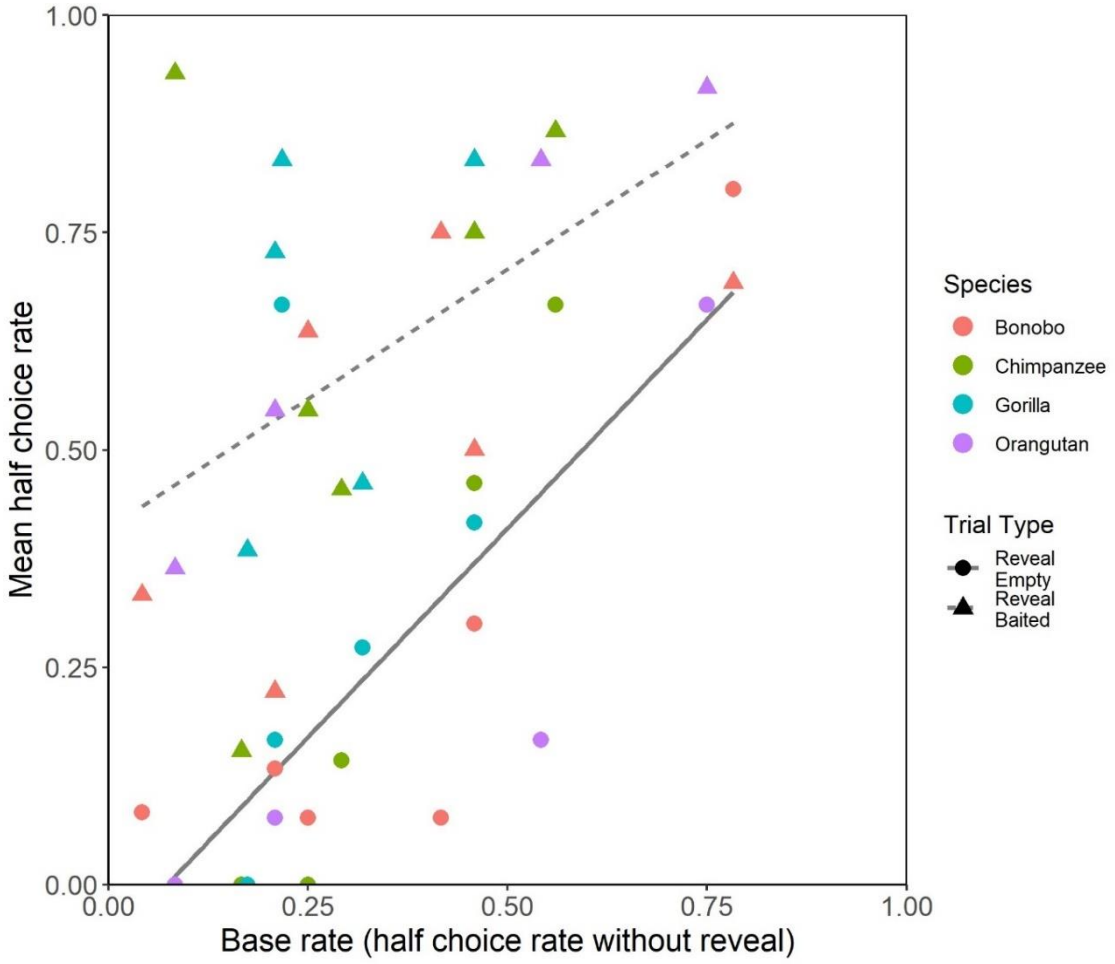


Figure 3.4.4: The relationship between taking the half-piece in Experiment 2 and subject's base rate from post-decision wagering without information (Chapter 2, experiment 2).

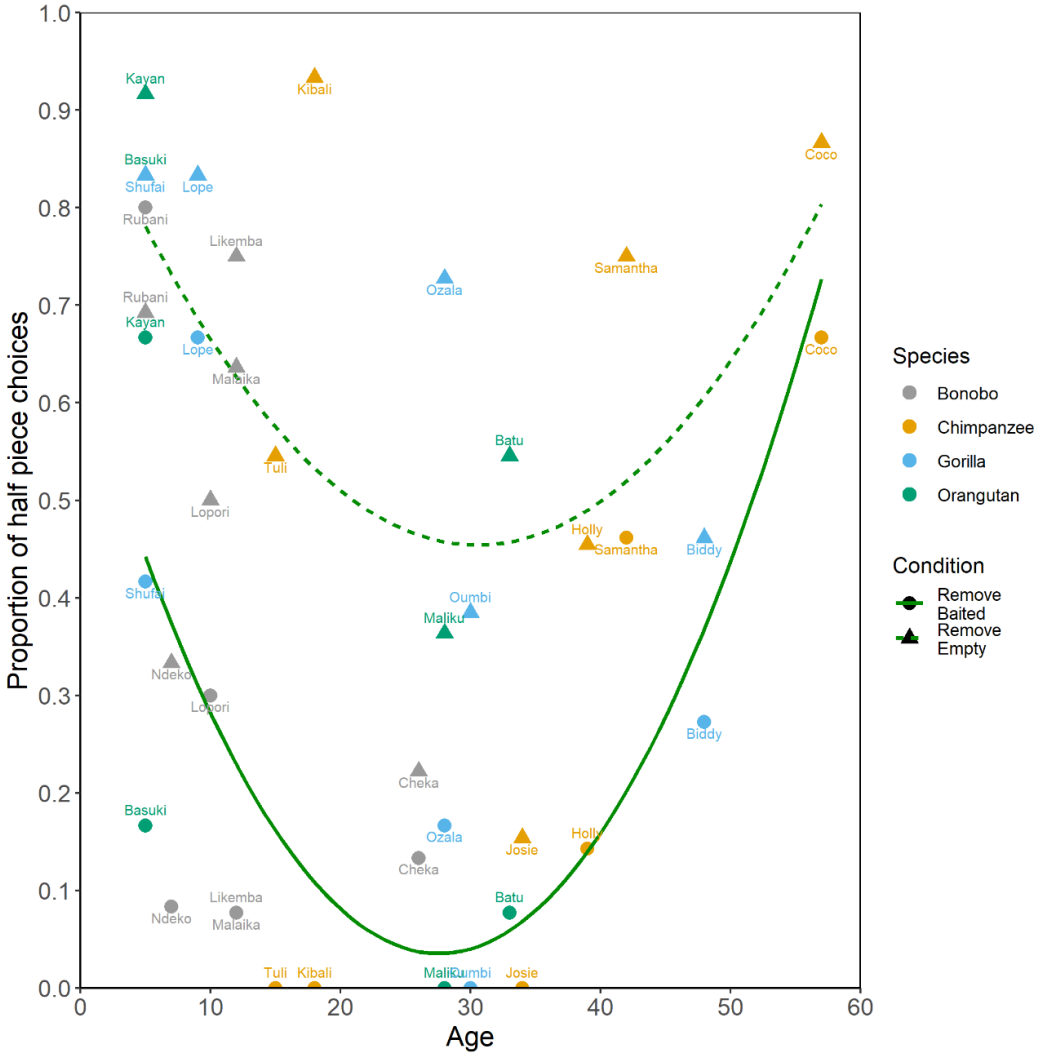


Figure 3.4.5: Individual half choice frequency by age and condition.

While we found no additional effect of age on inference ability, if we view the effect of age in isolation (Figure 3.4.5), we report a U-shaped relationship between inference ability and age. Call (2006) found a logarithmic relationship between age and inference by exclusion in the great apes, with performance improving up until around the age of 10 before levelling off. However, he did not test any subjects over the age of 32, so may have failed to capture age-related cognitive decline within his sample. Hopkins and colleagues (2021) tested 218 chimpanzees using the primate cognitive test battery (PCTB) (Herrmann et al., 2007), and found that the middle aged cohort (24-36 years) outperformed both their younger (< 26 years) and older (> 36 years) peers. Furthermore, when retested after a period of 1-7 years, the oldest individuals showed the greatest decline. Notably, the PCTB does not have tests for response inhibition, but in light of our results from Chapter 2, it is plausible that changes in this measure have played a contributing role in the results found in the literature.

However, when comparing the reveal empty trials with visible trials (Figure 3.4.1), the finding that apes took the half piece significantly more in the reveal-empty trials suggests that they are not treating their inferences as equivalent to visual evidence. This contrasts with the children tested by Mody and Carey (2016), who chose the target cup at equivalent rates in the 3- and 4-cup experiments. Notably, although the ~60% rate of taking the half piece in reveal baited trials is above the base rate from the previous experiment, it is also a long way from the 100% as expected by deductive inference. This provides support for the conclusion that apes are treating the decision probabilistically (Call, 2022; Engelmann, Haux, et al., 2023; Hanus & Call, 2014; Rescorla, 2009). Under this explanation, subjects need only trust a visual representation to a higher degree than an inferential one to produce this difference. While we did not find support for this hypothesis when we tested it in Chapter 1, the inference which subjects were required to make was simply object permanence, rather than inference by exclusion. Which, despite being reliant on the same spaciotemporal relations, is simpler as it does not require negation.

As one individual, Kibali, showed results comparable to Velu in Experiment 1, we can say that Velu's results are not exceptional and, as a naïve individual also performed to this level from their first session, we can conclude that they have not answered correctly based on associative learning. Nevertheless, it is possible that the apes are actually not approaching the problem inferentially and, in fact, the actions of the experimenter have primed these responses through stimulus enhancement. In the reveal baited condition, the experimenter reaches down to remove the whole piece before replacing it with the half piece, thus drawing the subject's attention to that position and making them more likely to select it. In contrast, in the reveal empty condition they did not do so, therefor it is possible that the mismatch between the behaviour of the experimenter in each condition is the actual cause of the reported effect.

3.5 Experiment 3

In Experiment 3 we modified the protocol of Experiments 1 and 2 to counter the possibility of solving the task via stimulus enhancement. Instead of placing the half-grape in the place of the removed cup, the experimenter placed it in a 3rd position, either to the left or the right of the pair, counterbalanced between trials. Additionally, in the *remove empty* condition the experimenter mimed the action of removing a grape to match their actions between conditions. Data were collected approximately 1 year after Experiment 1.

Methods

Participants

This experiment was conducted with the Edinburgh chimpanzees, 9 individuals took part in this experiment including a 41-year-old female (Sophie) who had not participated previously (Table 3.3.1)

Procedure

The baiting procedure followed that of experiments 1 and 2 except that the pair of cups were not placed centrally on the sliding table. The table was instead divided into 3 positions and the cups were placed into either positions 1 and 2 or 2 and 3, counterbalanced between trials. The procedure for the subject's first choice was the same as in previous experiments, but when revealing the contents of the unchosen cup the experimenter matched their actions between conditions by miming the removal of a grape in the reveal empty condition. Finally, to avoid stimulus enhancement, when placing the half grape, the experimenter did not place it in the position of the removed cup, but instead in the unoccupied 3rd position, then gave the subject their second choice. This experiment did not include any certain trials.

Two individuals (Frek and Paul) received apple pieces instead of grapes, the large and small pieces were 1/16th and 1/32nd of an apple, respectively. Two individuals, Sophie and Velu only completed one block within the 12 available sessions, all others completed two full blocks. Inter-coder reliability based on the first choice, removed cup contents and second choice for 15% of trials was excellent (kappa = .954, n = 32).

Results and Discussion

From 192 trials, 103 were *remove baited* and 89 *remove empty.* Figure 3.5.1 shows the mean rates of choosing the half piece as a function of removed cup contents. The difference between conditions was significant (t-test, t = -8.48, df = 8, p < .001) as was the difference from chance (*remove empty*: t-test, t = -8.18, df = 8, p < .001; *remove baited:* t-test, t = 4.77, df = 8, p = .001). Moreover, Fisher's exact tests showed that 6 of 9 individuals correctly adapted their choice behaviour based on the contents of the unchosen cup (Table 3.5.1). These results reinforce our data from experiment 2, that in 2-cup task chimpanzees can solve both variants of the disjunctive syllogism.

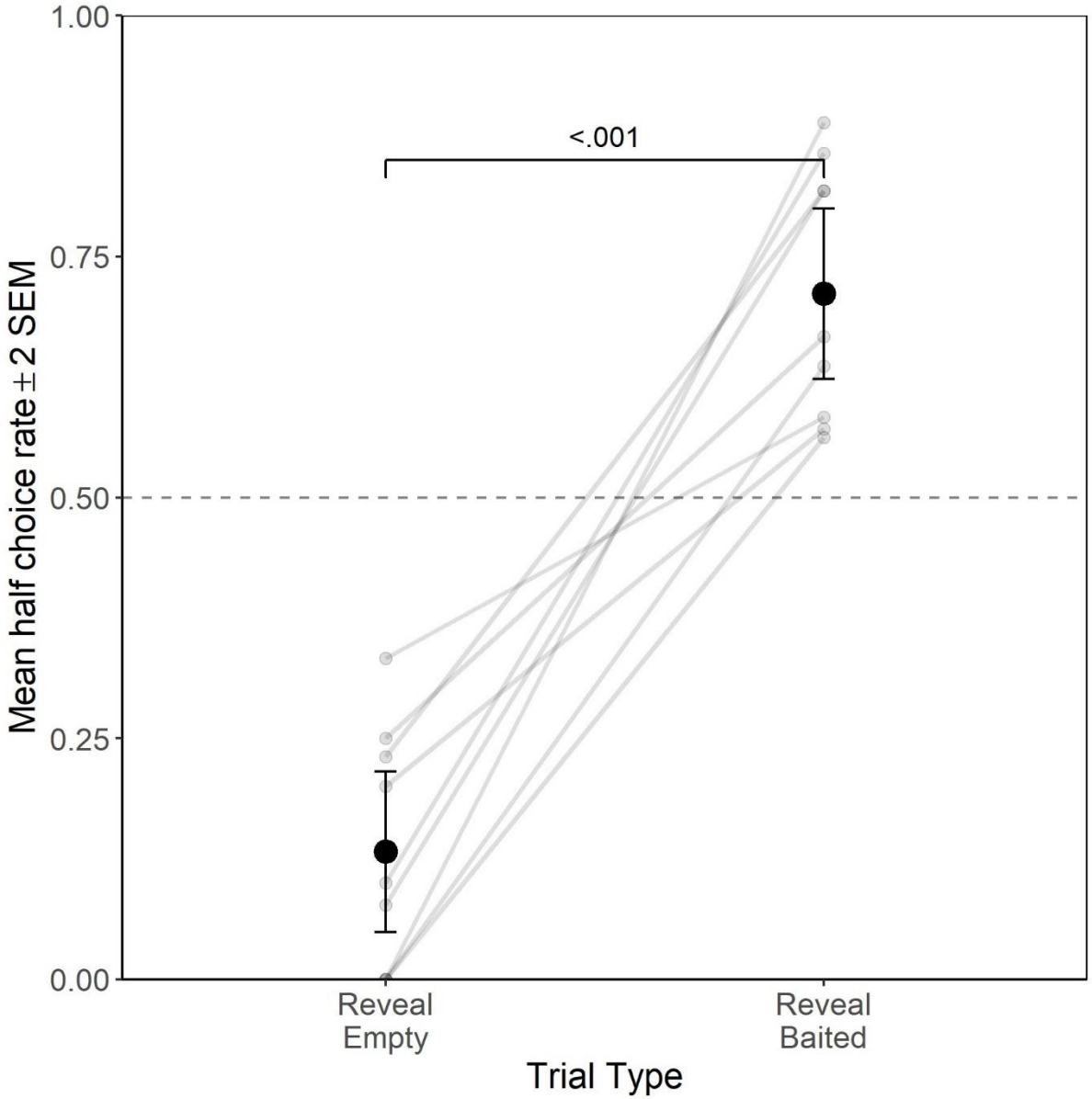


Figure 3.5.1: Half-choice rates by condition in experiment 3. The hashed line shows chance at 50%.

To test whether stimulus enhancement was playing a role in chimpanzee's performance we compared half-choice rates in trials where the half grape was on the same side as their original choice compared to trials where it was on the opposite side of the unchosen cup, but found no difference for either reveal empty (t-test,$t_{8,}$= 1.34 , p = 0.217) nor for reveal baited trials (t-test,$t_{8,}$= 1.75, p = 0.119), meaning that we can rule out this non-cognitive explanation.

Table 3.5.1: Individual half choice rates for Experiment 3. Fisher's exact test, with Holm-Bonferroni correction.

ID	Reveal empty	Reveal baited	Fisher Test (p)	Fisher Test (p.adj)
Edith	0	0.636	0.001	**0.003**
Eva	0	0.562	0.009	**0.021**
Frek	0.077	0.818	0.001	**0.002**
Kilimi	0.25	0.667	0.100	0.128
Lucy	0.333	0.583	0.414	0.414
Paul	0.231	0.818	0.012	**0.022**
Qafzeh	0.1	0.857	0.001	**0.002**
Sophie	0.2	0.571	0.293	0.330
Velu	0	0.889	0.018	**0.027**

Overall performance was significantly better in the modified version than the original (t-test, t = -3.78, df = 6, p = .009). However, the rates of choosing the half-piece were not different in either the *remove empty* (t-test, t = 1.55, df = 6, p = .173) nor the *remove baited* condition (t-test, t = -1.01, df = 6, p = .318). This suggests that the modified paradigm increased comprehension rather than biasing responses in one direction, potentially through increasing the salience of the half-choice as independent of the first choice. Fitting a GLMM as in Experiment 1, we still find a main effect of condition on half choice rates. However, we don't find an interaction between session and condition, ruling out a learning effect (Appendix 1, Table 8.2.2). The prolonged period between experiments 1 and 3, combined with a failure to find evidence of learning in either experiment would suggest that the results of Experiment 3 are not a learned association.

3.6 General Discussion

This set of experiments demonstrates that great apes can reason via the disjunctive syllogism when tested under a 2-cup search paradigm. We tested whether, when given information about the unchosen cup, subjects were able to infer the contents of their chosen cup and choose adaptively between it and a half-piece. From the first session, chimpanzees adaptively adjusted their half-choice rates in response to the contents of the unchosen cup. While the group unexpectedly regressed to chance in their second session, two individuals continued to choose adaptively throughout. In Experiment 2 we repeated the same paradigm but with a naïve sample comprising all 4 great ape species, finding that apes adaptively switched based on the contents of the unchosen cup and that there was no difference between species. Finally, in Experiment 3, we retested the Edinburgh chimpanzees while controlling for stimulus enhancement, finding

that under the modified paradigm those subjects who had passed previously continued to do so, but also that the group passed as a whole.

Significantly we have demonstrated that apes not only adapt the frequency with which they take the half piece in response to the contents of the revealed cup, but in both conditions they do so at a rate different from chance. This is true whether you take chance to be the 50% expected if subjects were picking randomly, or their individual rates when the unchosen cup wasn't revealed. This represents a unique result in the literature. The chimpanzees tested by Engelmann and colleagues (2022) were above chance in the reveal baited condition but at chance in reveal empty, while baboons showed the opposite response, being above chance in reveal empty trials but at chance in reveal baited (Ferrigno et al., 2021). Finally, while the chimpanzees tested by Call (2022) performed well in both conditions, in the reveal baited condition the measure of inference was indifference, so responding at chance was coded as correct. Therefore, ours is the first result that can demonstrate a full understanding of both variants of the disjunctive syllogism.

This study differed from the disjunctive syllogism tasks used by Engelmann (2022) and Ferrigno (2021) in that those tasks required subjects to track 2 food items and represent 4 possibly correct locations[18]. While chimpanzees have excellent memory for their own actions on search tasks (Völter et al., 2019; c.f. Read et al., 2022), it is plausible that working memory constraints have contributed to primate failures in the 4-cup variant. Crucially, this may not be a simple working memory deficiency, but rather a difficulty with maintaining concurrent mental models of the world, which has been suggested to dictate the upper limit of human reasoning capacities (Johnson-Laird, 2010). Alternatively, the constraint could be the requirement to actively inhibit searching for one of two identical food pieces. While not wholly explanatory in their own right, these could each be one of a number of factors, possibly additive in nature, resulting in cognitive load masking competence.

However, while reverting to a simplified task, a notable critique of 2-cup-1-item inference tasks is that they can be solved without understanding the concept of *therefore*, and instead marking the options independently as '*maybe-A, maybe-B* ', when A is shown empty the probability of B does not change and remains a '*maybe*' but is the only possible option. Performance in this experiment cannot be explained by this reductive explanation, as the data show that apes represent a dependency between the probabilities of A and B because, for multiple individuals,

[18] Call's (2022) 3-cup 1-item task also only used 1 food piece and found affirmative results in both conditions.

the frequency of a second behaviour (taking the fractional reward) was modulated by the contents of the revealed cup.

Mody and Carey (2016) argue that because children choose at comparable rates between a certain and an inference condition, that they are using logical reasoning. If we compare the ~10% frequency of taking the half-piece in reveal empty trials of Experiment 3 to the ~10% rate in visible trials of this paradigm (Experiments 1 & 3, Chapter 2), we can draw the same conclusion that Mody and Carey (2016) drew for children. However, it does raise the question as to why this relationship does not apply for the Twycross apes, for whom we directly interspersed visible trials to allow comparison. This could be down to several reasons. Firstly, the Edinburgh group's experience with the task in Experiment 1, this could be a purely associative response and the previous exposure to the task is what caused the Edinburgh chimpanzees to perform better in Experiment 3. However, an associative strategy would produce a pattern of initially random responses, before reaching ceiling and remaining there. This is the opposite of what we observed, performance decreased in the second session of experiment 1 and remained constant in experiment 3. Alternatively, it could be a task factor as suggested above, placing the half grape in a third position highlights it as separate from the first choice, while matching actions between conditions actually draws attention to the absence of the whole grape in the reveal empty condition.

Finally, there could be group differences between the Edinburgh and the Twycross groups which were masked by the performance of the Edinburgh group in the first experiment. While experiment 2 has showed that even a naïve group has the capacity for inferential reasoning, it could be that the confidence which they place in their inferences is lower due to a lack of experience. While the Twycross apes receive cognitive enrichment as part of their regular schedule at the zoo, they do not have experience with face-to-face cognitive testing.

The regression to chance in the second session of Experiment 1 is an unexpected result. Specifically, if there were to be differences between sessions, we would expect to see a learning effect, an improvement from the second session to the first. However, that is not the case. While attempting to explain away negative results should be a cautionary endeavour, during this testing period, the caregivers at Edinburgh Zoo were reintegrating a subgroup of chimpanzees who had been housed separately but within the same building for approximately 2 years. This meant moving to an artificial fission-fusion dynamic, in which some individuals would be moved back and forth from the subgroup and vice-versa before the groups were fully reintegrated. This resulted in understandably high levels of stress for the chimpanzees including but not limited to high rates of male-initiated aggression. As testing takes place within a

communal area and these tasks require both cognitive effort and concentration, it is reasonable to consider that this negatively impacted performance. Moreover, as subjects participated in testing on different days, this could have impacted different individual to different extents. In support of this, if we compare the performance of other researched-experienced ape groups tested under 2-cup paradigms (Call, 2006, 2022; Engelmann, Haux, et al., 2023) we see that the failure of the Edinburgh chimpanzees during the second session is an outlier.

However, this paradigm does not discern whether subjects are using full deductive inference or simply adjusting the *probability* that the remaining cup contains a grape and then choosing rationally between it and a half-grape (Rescorla, 2009). The 4-cup paradigm (Mody & Carey, 2016) also cannot rule out this explanation and future studies should explicitly aim to distinguish between inductive and deductive reasoning. Nevertheless, while we cannot declaratively state that subjects are considering certainty in a modal sense, we have demonstrated, that in a 2-cup 1-item task great apes appear to reason via both variants of the disjunctive syllogism and that language is not necessary for this ability to emerge.

4. The influence of methodology on the detection of disjunctive reasoning in great apes.

4.1 Abstract

Chimpanzees excel at inference tasks which require that they search for a single food item from partial information. Yet, when presented with 2-item tasks which test the same inference operation, chimpanzees show a consistent breakdown in performance. Here we test a diverse zoo-housed cohort (n = 24) comprising all 4 great ape species under the classic 4-cup 2-item task, previously administered to children and chimpanzees, and a modified paradigm administered to baboons. The aim of this study is to delineate whether the divergent results reported from the literature are taxonomic differences or artefacts of the paradigms, while extending the literature to cover the remaining great ape species. We find that apes adaptively adjust their choice behaviour in both variants of the paradigm, but that they perform better in trials where the information provided rules out a location rather than removes one of the food items. In a second experiment involving those subjects who passed the first, along with a group of naïve subjects, we test whether subjects were able to apply the logical operation selectively by including control trials where the correct response is reversed. Performance in standard trials breaks down with the addition of control trials, meaning that if apes did solve the first experiment logically, they are not capable of applying that logic flexibly. We then discuss whether the 4-cup paradigm is a suitable test of logical reasoning in great apes.

4.2 Introduction

To understand the significance of advanced reasoning to human existence a reader need only note the detail and enthusiasm with which errors in human reasoning have been documented (Kruger & Savitsky, 2004). Yet our capacity for advanced reasoning relies only on the appropriate combination of discrete inferences, new knowledge drawn only from held knowledge. As such, the extent to which these inferential processes dictate decision making in non-human species is a crucial element of comparative psychology and is the subject of intense debate.

In the language of logic, inference takes the form of a syllogism - two statements, or *premises*, which flow to a natural conclusion. For example, "dogs are mammals, all mammals have fur, therefor dogs have fur". When one premise contains two possibilities connected by the logical operator *or* it becomes a disjunctive syllogism – if either A or B is true, and A is false then B must be true. The disjunctive syllogism can more simply be referred to as inference by exclusion.

Inference by exclusion can be tested in the absence of language using a visual search task, choice behaviour in line with rational thought can be considered evidence of logical reasoning. To illustrate this Call's (2004) 2-cup 1-item task presents a subject, in his case great apes, with two upturned opaque cups one of which contains a target item. When participants were given indirect evidence by the experimenter shaking one cup or showing them the empty cup, they were able to infer the location of the target item. This paradigm has been used to illustrate inference by exclusion in primates (Call, 2006; De Petrillo & Rosati, 2020; Heimbauer et al., 2019; Petit et al., 2015), birds (Mikolasch et al., 2012; O'Hara et al., 2015, 2016; Pepperberg et al., 2013), domestic dogs (Erdőhegyi et al., 2007) and children as young as 23 months of age (Mody & Carey, 2016).

Two-cup tasks can, however, be solved by using a strategy of '*maybe-A maybe-B*', in which subjects need only mark each option as possible locations for the target item, then by avoiding the cup which they know to be empty. In its simplest format, the subject does not even need to mark either option as possibilities and can respond appropriately by choosing what is remaining after avoiding the empty cup. To test this non-inferential strategy, Call (2022) followed up the original experiment with a paradigm to explicitly test the *'avoid the empty cup'* hypothesis and found no support for it, suggesting that apes are in fact capable of using inference to solve these tasks.

However, Leahy and Carey (2020) challenge the notion of inferential reasoning in non-human animals by arguing that solving these tasks does not even require a full understanding of possibility. They propose that pre-verbal children and non-human animals possess only a minimal model of possibility, with learning the language of possibility being a necessary stage in scaffolding a mature understanding. Leahy and Carey's thesis is that human infants and primates lack the modal concepts *possible, impossible*, and *necessary*, with a minimal agent's model of possibility relying on a process of making only a single simulation of the state of the world which she will treat as fact. In the context of the inference by exclusion paradigm, a participant is not compelled to treat both options as possibilities. She can instead make a guess

and once she receives either confirmatory or conflicting evidence, keep or revise her decision accordingly.

The full- and minimal model of possibility can be delineated from one another by Mody and Carey's (2016) 4-cup task. The task involves an experimenter separately baiting behind a visual barrier two pairs of cups with one sticker each, such that each cup has a 50% chance of containing a sticker. They then remove one cup, show the child that it's empty, and give them a choice between the remaining three cups. Three-year-old children[19], but not younger, consistently pick the remaining cup within the pair. The authors cite this as evidence that they are able to reason through the disjunctive syllogism: A *or* B, *not* A *therefore* B.

Younger children and chimpanzees, instead pick the target cup 50% of the time, which Leahy and Carey (2020) interpret as evidence of them having made concrete guesses as to where the two stickers are, the first of which was incorrect, so they revised it, and then choose randomly between their revised guesses. Similarly, in a 3-cup 2-item task in which one cup represents certainty, apes and young children only choose it 50% of the time (Engelmann, Haux, et al., 2023; Hanus & Call, 2014; Leahy et al., 2022). Thus, irrespective of their inferential capacities, apes and young children appear to not give special status to a certain outcome.

While the original 4-cup paradigm reveals an empty cup to test the inclusive *OR*, *the sticker must be under A or B*, Gautam, Suddendorf and Redshaw (2021) extended the paradigm to include 'reveal-baited*'* trials to test the exclusive *OR* relation, *the sticker cannot be under A and also under B*. In their procedure, a sock puppet took the first guess and managed to find a sticker on 50% of trials. In these trials the correct choice is to switch to the alternate pair of cups, which children don't reliably do until five years of age[20]. Crucially the subject will only respond appropriately to the 4-cup paradigm if they represent the likelihood of each cup within a pair containing a sticker as being dependent on the contents of the other cup.

In an attempt to make the primate variant of the 4-cup task more closely resemble a natural foraging exercise, Ferrigno, Huang and Cantlon (2020) gave their subjects two choices, first from the initial 4 cups and then from the remaining 3, to create a natural experiment whereby 50% of trials would be reveal empty and 50% would be reveal baited. This new 2-choice paradigm takes the form of a logically directed sequential search, rather than a response to an

---

[19] Later revised to 2½ by Gautam, Redshaw and Suddendorf (2021a).
[20] The minimal model predicts 100% accuracy in the reveal-baited condition, when one of the two pieces are removed the subject revises their model with regard to the target cup, and searches for the remaining sticker in the alternate pair.

experimenter manipulation. The authors tested nine zoo-housed olive baboons (*Papio anubis*), of which 4 passed the pre-tests. As a group, these 4 baboons scored above chance in reveal empty trials, 3 subjects were individually above chance and 2 were above the 50% chance rate as set out by the minimal model of possibility. However, the 3 baboons who passed the reveal empty trials were precisely at chance in the reveal baited trials. This demonstrates that reasoning via the inclusive disjunction in a 4-cup paradigm is not limited to humans while supplementing the developmental evidence proposing that reasoning via the exclusive disjunction is more complex than the inclusive disjunction.

Engelmann et al (2022) have subsequently tested chimpanzees under the 2-, 3- and standard 4-cup paradigms, with both the reveal empty and the reveal baited trial-types. Consistent with the literature, they found that a majority of apes passed the 2-cup task but were at 50% in the 3-cup task. However, in contrast to young children (Gautam et al., 2021b) and to the baboons tested by Ferrigno et al. (2020), the chimpanzees correctly stayed within the pair on just 50% of reveal empty trials, but fared better on reveal baited trials, correctly switching to the alternate pair on ~85% of trials. While the difference between children and chimpanzees is unclouded, making claims regarding differences between chimpanzees and baboons is problematic because of the differences between the paradigms, and is something which Englemann and colleagues (2022) refrained from doing. Here we can test great apes under both paradigms, to ascertain to what extent these differences are artefacts of their respective paradigms; and secondly, as in Chapter 2, we can investigate whether the findings of a research experienced chimpanzee cohort translate to a naïve group comprised of all 4 great ape species.

The current research.

The current study is motivated by two missing pieces of data from the literature. First, it is unknown how chimpanzees would respond to the two-choice version of the 4-cup task that Ferrigno et al. (2020) used with baboons. One of the reasons that these authors mentioned for designing this task was to make it more similar to a natural foraging task, which might reduce the cognitive load of the original task because subjects directly searching and uncovering the first cup might be more memorable than observing an experimenter uncovering it. Second, it is unknown how great apes other than chimpanzees fare in the 4-cup task. Such data are important to contribute to elucidate the taxonomic distribution of cognitive abilities.

The current research aims to tests the 4- great ape species under both the standard 4-cup task (Mody & Carey, 2016, here referred as the one-choice condition), and the 2-choice variant (Ferrigno et al., 2020, here referred as the two-choice condition). Subjects receive a block of each variant, counterbalanced by order. The two tasks test whether subjects can adaptively

switch between pairs based on the contents of the first revealed cup. The difference between these variants is that the information is either from an experimenter's manipulation or their own directed search. As they are both testing the same underlying ability, we hypothesise there to be a correlation between performance on the two tasks. In experiment 1 we tested a group of relatively inexperienced great apes under both 4-cup paradigms, to ascertain to what extent the differences within the literature are artefacts of their respective paradigms; whether a naïve group will replicate the results of the experienced group tested by Engelmann et al (2022); whether non-chimpanzee great apes will perform like chimpanzees, and whether their performance in our 2-cup disjunctive syllogism task (Chapter 3) translates into the 4-cup variant. In experiment 2 we then conduct a follow-up experiment to rule out non-inferential strategies and test whether those subjects who passed the 4-cup task can flexibly apply the same reasoning to a modified paradigm.

4.3 Experiment 1

Methods

Participants

We tested 24 apes (2 orangutans, *Pongo pygmaeus*, 5 gorillas, *Gorilla gorilla*, 7 chimpanzees, *Pan troglodytes,* and 9 bonobos, *Pan paniscus*) housed at Twycross Zoo, England. Four apes failed the pre-test (1 chimpanzee, 1 bonobo and 2 gorillas), 1 individual passed the pre-test but only completed 7 test trials within the available sessions, so her data were not included. The sample for experiment 1 comprised 19 apes, 2 orangutans, 3 gorillas, 6 chimpanzees and 8 bonobos (11 female, mean age = 19.5 years). Detailed demographic data and rearing history can be found in Table 4.3.1. The apes lived in natural groups and had access to both indoor and outdoor space with vegetation. Water was available *ad libitum* during testing and the apes received regular feedings of a wide variety of fruits and vegetables throughout the day, and additionally received further enrichment. Testing took place within a communal area and was completely voluntary. The apes were largely inexperienced with cognitive testing, the experiments outlined in chapters 2 and 3 were the orangutans' and gorillas' first experience with cup-based search tasks, while the chimpanzees and bonobos participated in that task and one other.

Table 4.3.1: Demographic details of subjects, age is measured in whole years at the date of their first session.

Species	ID	Sex	Age	Rearing History	Conditions
Bonobo	Cheka	Female	26	Parent	Both
	Likemba	Female	12	Unknown	Both
	Lola	Female	3	Parent	Both
	Lopori	Female	10	Hand	Both
	Lucuma	Male	19	Parent	Both
	Malaika	Female	12	Parent	Both
	Ndeko	Male	7	Parent	Both
	Rubani	Male	6	Unknown	Both
Chimpanzee	Flyn	Male	36	Hand	Two-choice
	Holly	Female	39	Parent	Both
	Josie	Female	34	Hand	One-choice
	Kibali	Male	18	Parent	Both
	Tuli	Female	15	Parent	Both
	Victoria	Female	32	Hand	One-choice
Gorilla	Lope	Male	9	Parent	Both
	Ozala	Female	28	Parent	Both
	Shufai	Male	5	Parent	Both
Orangutan	Batu	Male	33	Parent	Both
	Kayan	Female	5	Parent	Both

Apparatus

The experimenter (E) sat opposite the subject with a sliding table (630mm x 300mm) placed between them. Two matching pairs of coloured cups (Ø = 82 mm, height = 120 mm) were used as hiding locations, spaced equidistantly (~60mm) at the front of a sliding table. Cups were arranged in sets, so positions 1 and 2 made up set 1 and were one colour, and positions 3 and 4 made up set 2 and were a different colour. The sets were either red and blue (orangutans and chimpanzees) or pink and yellow (bonobos and gorillas). A U-shaped occluder (HxWxD: 300mm x 380mm x 150mm) was used to block the subject's view of the cups during baiting. Cubes of raw sweet potato ($1cm^3$) were used as the target items.

Procedure

Pre-test

The experimenter showed the subject a sweet potato piece and visibly placed it into one of the two sets of cups, visiting both cups within the pair in a left-right direction and depositing it under one of them. The E then showed the subject that their hands were empty before repeating the procedure with the second sweet potato piece and the alternate set. They then lifted one of the cups and moved it to the back of the table and, if it had been baited, discarded the food

piece, before sliding the table to its forward position so that the subject could reach through the mesh and select one of the three remaining cups by touching it. To pass the pretest subjects had to select a correct cup (chance = 0.5) on 8 from 10 trials within a maximum of 2 sessions[21].

Test

Test trials were split into one-choice and two-choice conditions. Two chimpanzees only completed the *one-choice* condition, and 1 chimpanzee only completed the *two-choice* condition. All other subjects received 4 consecutive sessions of each condition, each consisting of 8 trials for a total of 32 trials per condition. Block order was randomised within species, with half of the subjects receiving the one-choice condition first, and half receiving it second.

The *one-choice* condition followed the protocol of Gautam, Suddendorf and Redshaw (2021), but with the experimenter removing the first cup in place of the sock puppet. The baiting procedure was the same as the pre-test but the baiting took place behind an occluder and cups started at the back of the board. To start the trial, E lifted the cups and placed them at the front of the board (to demonstrate that they were empty), then placed the occluder in front of one pair. E then showed the subject the food item above the occluder before hiding it under one of the two occluded cups, as in the pre-test they lifted each cup sequentially in a left-right direction. E then lifted the occluder and placed it down in front of the second pair of cups and repeated the baiting procedure with a second food item. The order which E baited the pairs was counterbalanced within each session. E then lifted one of the cups and moved it to the back of the board and, and, if baited, discarded its contents into a bucket on the floor, before sliding the table to the subject to choose from the remaining three cups. The two possible trial types were *reveal empty,* where the removed cup was empty, and *reveal baited*, where the removed cup had been baited (before E discarded the food item). Subjects received 16 trials per condition, spread equally across the four sessions, the location of the food items, the identity of the removed cup (locations 1-4) and its status (baited/empty) were counterbalanced between trials.

The protocol of the *two-choice* condition followed Ferrigno, Huang and Cantlon's (2021) procedure. The baiting procedure was identical to the one-choice condition, but instead of the experimenter removing a cup, the subject was given two choices. If the subject could not detect the location of the target item on their first choice, they should guess correctly on half of the trials (which they did, see below). For comparison with the one-choice condition, trials where

[21] Chance is set to 0.5 because for 50% of trials a baited cup was removed, therefore one baited cup remained so p(Correct) = 0.33 and for 50% an un-baited cup was removed, so two baited cups remained and p(Correct) = 0.67.

subjects were correct on their first guess will be referred to as *reveal baited*, and those where they were incorrect will be *reveal empty*.

Coding and Data Analysis

As the cups were arranged in sets, all analysis is based on the rate of switching between sets, so if cup 3 was removed or chosen in the first guess then choosing either 1 or 2 would be coded as switching, while choosing 4 would not. All sessions were videotaped, and the subjects' choice was live coded by the experimenter. A second coder blind to the purpose of the experiment coded 15% of trials from the video footage (inter-coder reliability was excellent, Cohen's Kappa = 0.978). All statistics are paired tests unless otherwise stated.

Results and Discussion

Figure 4.3.1 shows how the rate of switching sets varied as a function of variant and trial-type. To test for reasoning in line with the disjunctive syllogism, we compared frequencies of switching sets between reveal empty and reveal baited trials. To do so we fitted a binary logistic regression with a logit link function, using the fixed effects of trial type (reveal empty/ reveal baited), variant (one-choice/two-choice), species and trial number (1-48), and the random effect of ID. As the difference in switch rates between reveal empty and reveal baited trials is our measure of inference, we also included the two-way interactions between revealed cup contents and each of the other predictors. The model with the random effect was not an improvement over a GLM with only the fixed effects ($\chi^2 = 2.712$, df = 1, p = .100), so we continued with the GLM. This model fitted the data better than a model without the interactions ($\chi^2 = 17.612$, df =5, p = .003), and a null model ($\chi^2 = 64.0$, df = 11, p = < .001). Coefficients from the final model can be found in Table 4.3.2.

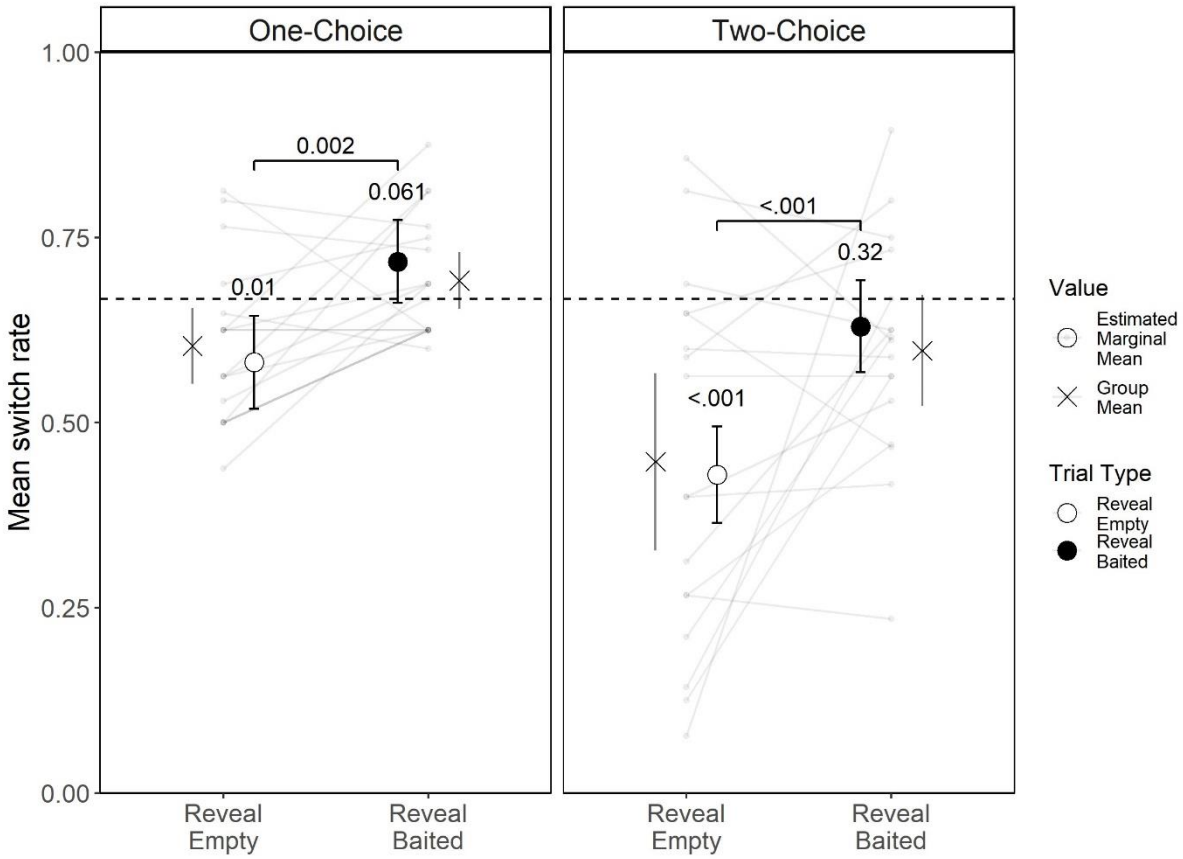


Figure 4.3.1 Group and individual level rates of switching pairs by trial type and condition. Points show the estimated marginal means from the model (averaged across species'), along with paired contrasts for switch rates between trial types and z-tests against chance (null = 0.667). The hashed line represents responding at chance, crosses show group means

We find support for a main effect of revealed cup contents (χ^2 = 15.646, p < .001) and no difference in the size of the effect between the one-choice and two-choice conditions (χ^2 = 0.735, p = .391), this shows us that apes can adjust their choice behaviour adaptively in both a self-directed sequential search and in response to an experimenter manipulation. Secondly, we find no support for an interaction between cup-contents and trial number (χ^2 = 1.891, p =.169), so we can reject the conclusion that this is a learned response.

We do report a difference in effect size by species (χ2 = 14.39, p = .002)(Figure 4.3.2). Pairwise contrasts reveal a significantly higher rate of incorrectly switching pairs in reveal empty trials for chimpanzees compared to all other species (Table 4.3.3). The inclusion of the 3-way interaction between trial type, condition and species did not improve the fit of the model to the data (χ2 = 9.33, p = .156), suggesting these species differences are consistent across conditions. While the chimpanzees did not adapt their switch rate in response to the contents of the revealed cup (β = -0.396, CI95 (-1.083, 0.292)), this is not necessarily reason to believe that this response

is indicative of the species, as similar studies have demonstrated inferential reasoning in chimpanzees under comparable paradigms (Call, 2022; Engelmann, Haux, et al., 2023).

*Table 4.3.2: Coefficients from a model to predict the binary outcome of switching given condition, trial type, species, and trial number. (switch ~ trial-type*condition + trial-type*species + trial-type*trial-number. Family = binomial(link = "logit"))*

	Estimate	CI (2.5%)	CI (97.5 %)	p-value
(Intercept)	1.213	0.729	1.697	<.001
(species)Bonobo	-0.715	-1.146	-0.285	.001
(species)Gorilla	-0.782	-1.314	-0.251	.004
(species)Orangutan	-1.323	-1.959	-0.688	<.001
(Trial type) Reveal Baited	-0.396	-1.083	0.292	.259
(condition) Two-choice	-0.612	-0.955	-0.269	<.001
Trial number (1-48)	-0.008	-0.021	0.005	.217
(species)Bonobo: (Trial type) Reveal Baited	0.409	-0.202	1.020	.189
(species)Gorilla: (Trial type) Reveal Baited	0.755	-0.016	1.527	.055
(species)Orangutan: (Trial type) Reveal Baited	1.706	0.792	2.620	<.001
(Trial type) Reveal Baited: (condition) Two-choice	0.213	-0.278	0.704	.396
(Trial type) Reveal Baited: Trial number	0.013	-0.006	0.031	.174

Table 4.3.3 Pairwise contrasts for differences in switch rates between species by condition in Experiment 1. Bold results show significant differences at an alpha level of .05.

Trial Type	Contrast	Estimate	CI (2.5%)	CI (97.5 %)	p-value
Reveal Empty	Chimpanzee - Bonobo	0.715	0.285	1.146	**.006**
	Chimpanzee - Orangutan	1.323	0.688	1.959	**< .001**
	Chimpanzee - Gorilla	0.782	0.251	1.314	**.021**
	Bonobo - Orangutan	0.608	0.023	1.193	.174
	Bonobo - Gorilla	0.067	-0.404	0.538	.992
	Orangutan - Gorilla	-0.541	-1.204	0.122	.380
Reveal Baited	Chimpanzee - Bonobo	0.306	-0.128	0.740	.510
	Chimpanzee - Orangutan	-0.383	-1.040	0.274	.663
	Chimpanzee - Gorilla	0.027	-0.532	0.586	1.000
	Bonobo - Orangutan	-0.689	-1.296	-0.082	.117
	Bonobo - Gorilla	-0.279	-0.780	0.222	.694
	Orangutan - Gorilla	0.410	-0.293	1.112	.663

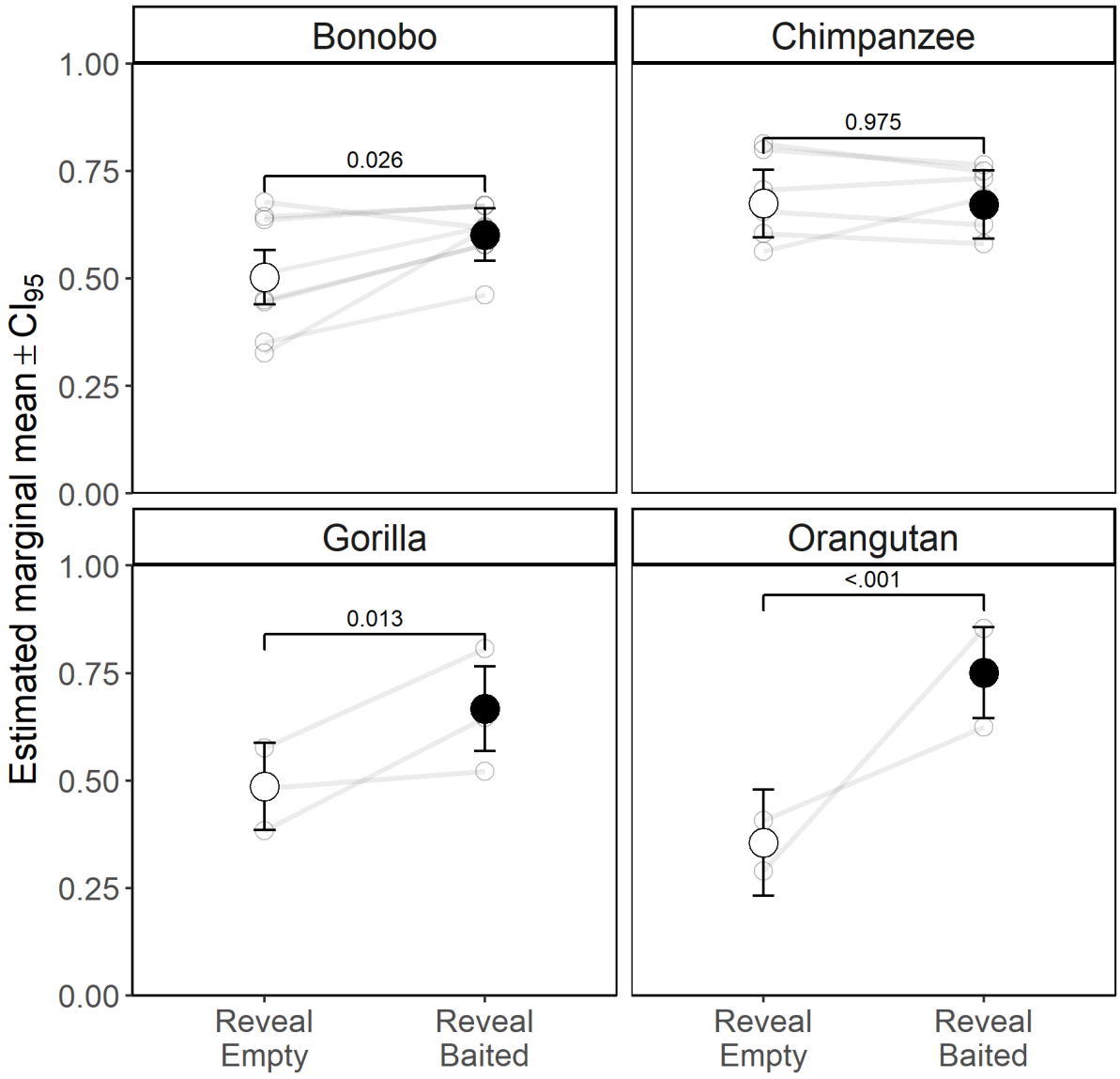


Figure 4.3.2 Estimated marginal means from a model to predict switching in Experiment 1 by trial type and species. Light grey points show individual level switch rates.

Table 4.3.4 Mean switch rate and SEM for each species by trial type and condition

	One-Choice		Two-Choice	
	Reveal Empty	Reveal Baited	Reveal Empty	Reveal Baited
Bonobo (n = 8)	0.59 (0.04)	0.69 (0.03)	0.42 (0.09)	0.51 (0.04)
Chimpanzee (n = 6)	0.68 (0.04)	0.68 (0.03)	0.68 (0.05)	0.67 (0.04)
Gorilla (n = 3)	0.58 (0.02)	0.69 (0.06)	0.38 (0.13)	0.63 (0.11)
Orangutan (n = 2)	0.5 (0.00)	0.72 (0.09)	0.19 (0.12)	0.76 (0.13)

We also find a main effect of condition, ($\chi 2 = 16.99$, df = 1, $p < .001$), showing that subjects are less likely to switch to the alternate pair in the two-choice paradigm where they engage in a sequential search rather than respond to an experimenter manipulation (Figure 4.3.1 Group and individual level rates of switching pairs by trial type and condition. Points show the estimated marginal means from the model (averaged across species'), along with paired contrasts for switch rates between trial types and z-tests against chance (null = 0.667). The hashed line represents responding at chance, crosses show group means). This overall effect of condition on switch rates is relevant because great apes have been shown to struggle with inhibiting the search of adjacent containers (Barth & Call, 2006; Mallavarapu et al., 2014). It is possible that the two-choice variant artificially inflates performance on the reveal empty trials (and conversely, primes reduced performance in the reveal baited trials). A hypothetical agent who chose randomly on their first guess and then chose an adjacent cup would only switch on 25% of trials. While this extreme case is not what our data show, and does not explain the difference in switch rate between trial types, it does illustrate how a policy of adjacent search could improve performance in reveal empty trials at the expense of reveal baited, highlighting the importance of testing for differences between conditions rather than simply absolute rates against chance.

When we do compare against chance (Figure 4.3.1), subjects switched at a rate lower than chance in reveal empty trials of both conditions (one-choice, z-test, z = -2.57, p = .010; two-choice, z-test, z = -7.02, $p < .001$) (one-choice, t-test, t17 = -2.46, p = .025; two-choice, t-test, t16 = -3.67, p = .002) but at chance in reveal baited trials (one-choice, z-test, Z = 1.88, p = .061; two-choice z-test, Z = -.99, p = .320)(one-choice, Wilcoxon, V17 = 101, p = .511; two-choice t-test, t17 = -1.85, p = .082).Which suggests that, like baboons and children between the ages of 2½ and 5, subjects are treating the *or* relation as inclusive (A or B, not A therefore B), but not exclusive (A or B, A therefore not B) (Gautam et al., 2021a). However, this comparison is limited because performance in reveal empty trials is notably worse than even the youngest age group tested, 2½-year-olds, who only switched pairs on 28% of reveal empty trials (Gautam et al., 2021b).

Our data show a key divergence from another group of chimpanzees tested under the one-choice 4-cup paradigm (Engelmann et al., 2022), who picked the target cup in 50% of reveal empty trials, while performing at close to ceiling (85%) in the reveal baited trials. While we report switch frequencies of 60% and 69% for reveal empty and reveal baited trials, respectively. Engelmann et al. (2022) proposed an alternative non-deductive theory to explain the results they found. Under this theory the subject need not engage in any simulation at all and instead can

simply mark two broad locations, one covering each pair, and reason that there is one item in each. In reveal baited trials one location is ruled out when the food is removed from it, so the subject chooses the other location, explaining the close to ceiling performance. In reveal empty trials, one cup is ruled out, shrinking that location but not ruling it out, and the subject chooses between the two locations, resulting in close to 50%.

While the authors do rule out that explanation on account of its failure to explain performance in alternate conditions of the original 3-cup task (Hanus & Call, 2014), it is significant that we see a divergence between groups of the same species. These conflicting results may be a consequence of task factors. For example, while they are essential for ruling out non-cognitive explanations, the presence and format of control trials also makes it more challenging to compare between studies. Engelmann uses control trials in which the cups are baited as a quartet rather than 2 pairs, which elegantly controls for ape's difficulty in inhibiting searching adjacent cups (Barth and Call 2006), but also gives the subjects additional experience with the baiting procedure. The authors did find a small effect of block order (control-test vs test-control), while it is highly likely that this simply an artefact of random assignment of the 13 individuals to one order or the other, there is the possibility that some element of the control procedure augmented performance in the test trials.

The conclusion reached by Engelmann and colleagues (2023) and by Hanus and Call (2014) is that apes reason probabilistically. Meaning that, without language, apes do not have access to a concept of certainty so cannot reason 'logically' but do however represent the contents of each cup as mutually dependent on each other, so inferentially update their predictions in light of new evidence. Notably, the 50% and 85% switch rates reported for the two trial types (Engelmann, Haux, et al., 2023) represent approximately equal adaptive divergences away from the chance rate of 66%. If we inspect individual trends in the left pane of Figure 4.3.1, we see a small but consistent difference between conditions, so in this sense, our data from the one-choice condition do conform to the literature, just to a lesser degree. It may be that this difference is due to differing levels of experience with object search paradigms, or cognition research more generally. It is also possible that the difference can be accounted for by Engelmann et al. (2022) only progressing those individuals who had chosen the target cup above chance in the 2-cup task onto the 3- and 4-cup tasks, thus screening out individuals for whom the inference operation was the limiting factor.

We observed a large amount of individual variation but the data did not indicate a correlation between performance in the two variants ($r_{16} = 0.05$, $p = .841$), suggesting that the two tasks are not capturing the same underlying capacity. Figure 4.3.3 shows the individual switch rates by

condition and trial-type. We used Fisher tests to test for a contingency between switching and the contents of the revealed cup (Table 4.3.5). After accounting for multiple comparisons (Holm-Bonferroni), two individuals showed such a contingency (Kayan, $p < .001$, and Lope $p = .038$), but only in the two-choice condition.

Under the minimal model (Leahy & Carey, 2020), a minimal agent will switch pairs on 50% of reveal empty trials, when we set chance at this level we see that both of these individuals outperform a minimal agent (binomial test, one-tailed, chance = .50, Kayan: $p = .002$, Lope: $p = .006$), thus we can reject the minimal model as an explanation for their choice behaviour. Additionally, Kayan was also above chance in *reveal baited* trials (binomial test, one-tailed, chance = .66, Kayan: $p = .021$, Lope: $p = .385$), which suggests that, unlike the baboons tested under the two-choice paradigm (Ferrigno et al., 2021), she was treating the OR relation in both its inclusive and exclusive sense.

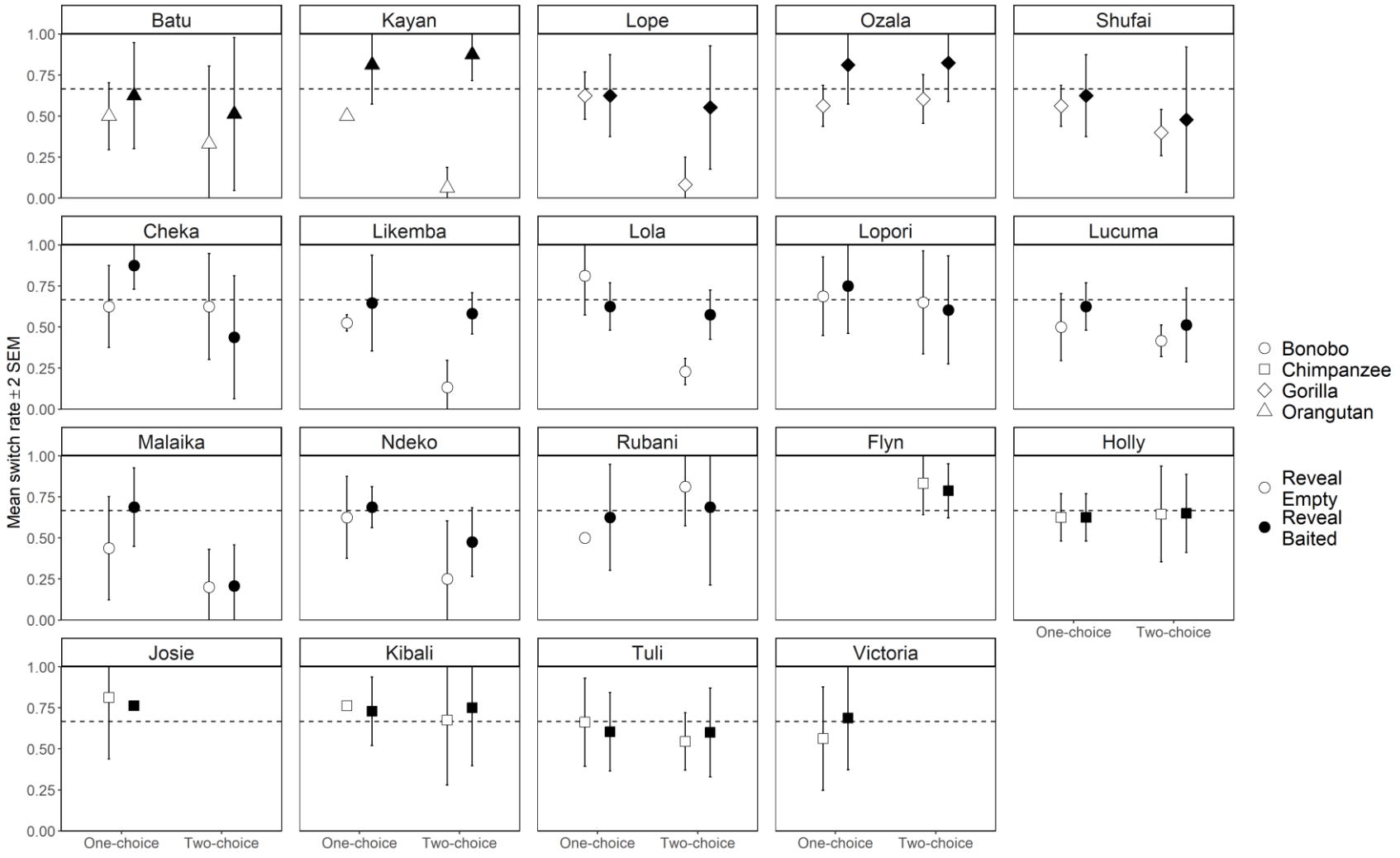


Figure 4.3.3: Individual switch rates by trial type and condition.

Table 4.3.5: Individual switch rates by condition and trial type and the p- value of a fisher's exact test for differences between trial types adusted for multiple comparisons using a Holm-Bonferroni correction.

		One-choice				Two-choice			
Species	ID	Reveal Empty	Reveal Baited	Fisher Test (p)	Fisher Test (p.adj)	Reveal Empty	Reveal Baited	Fisher Test (p)	Fisher Test (p.adj)
Bonobo	Cheka	0.625	0.875	0.22	1	0.647	0.467	0.476	0.854
	Likemba	0.529	0.667	0.491	1	0.125	0.562	**0.023**	0.127
	Lola	0.812	0.625	0.433	1	0.211	0.615	**0.03**	0.127
	Lopori	0.688	0.75	1	1	0.6	0.588	1	1
	Lucuma	0.5	0.625	0.722	1	0.4	0.529	0.502	0.854
	Malaika	0.438	0.688	0.285	1	0.267	0.235	1	1
	Ndeko	0.625	0.688	1	1	0.267	0.471	0.291	0.618
	Rubani	0.5	0.625	0.722	1	0.857	0.611	0.235	0.618
Chimpanzee	Flyn					0.812	0.75	1	1
	Holly	0.625	0.625	1	1	0.688	0.625	1	1
	Josie	0.8	0.765	1	1				
	Kibali	0.765	0.733	1	1	0.647	0.733	0.712	1
	Tuli	0.647	0.6	1	1	0.562	0.562	1	1
	Victoria	0.562	0.688	0.716	1				
Gorilla	Lope	0.625	0.625	1	1	0.143	0.667	**0.005**	**0.038**
	Ozala	0.562	0.812	0.252	1	0.588	0.8	0.265	0.618
	Shufai	0.562	0.625	1	1	0.4	0.417	1	1
Orangutan	Batu	0.5	0.625	0.722	1	0.312	0.625	0.156	0.529
	Kayan	0.5	0.812	0.135	1	0.077	0.895	**<.001**	**<.001**

Individual differences in cognitive performance have been extensively demonstrated within captive primates (e.g. Banerjee et al., 2009; Beran & Hopkins, 2018; Fichtel et al., 2020). Herrmann and Call (2012) reanalysed data collected as part of the primate cognitive test battery (PCTB) to look for individual differences among the 108 sanctuary living chimpanzees tested, the authors found that one individual's standardised score across the battery was $z = 2.12$, demonstrating that she was exceptional among her peers. The authors expand upon a point made in their initial PCTB study (Herrmann et al., 2007), that instead of a primate general intelligence, we should instead look at clusters of related intelligences, they analyse PCTB data from the apes at Leipzig Zoo ($n = 15$) using dimension reduction techniques and find that 3 factors can explain 82% of the variation in the data. The loadings of these 3 factors suggest distinct groupings for associative- (shape, colour and space tasks) and inferential processes (exclusion and causality), along with a third factor that loads tools, quantity and size. This suggests that, within the great apes, there is a clear variation along a spectrum of inferential ability, which provides the basis for the individual differences our data have revealed.

The demonstration that one individual can solve both variants of a 4-cup task is an exception within the literature and provides further evidence against language being an essential pre-requisite for reasoning via the disjunctive syllogism. However, it may be that this individual has happened upon a non-inferential strategy. If this were the case, we would initially expect to see performance being at chance before abruptly improving and remaining at ceiling for the remainder of the experiment. Figure 4.3.4 shows how switch rate varied throughout the 4 two-choice sessions for each subject who switched at a rate different from chance in at least 1 condition (one-tailed binomial test, $p = .05$). If we look at Kayan, and to a lesser extent at Likemba and Lola, we see that from the first session they are switching adaptively based on the contents of their first guess, and performance does not markedly improve. Lope, however, shows a profile that is more indicative of associative learning.

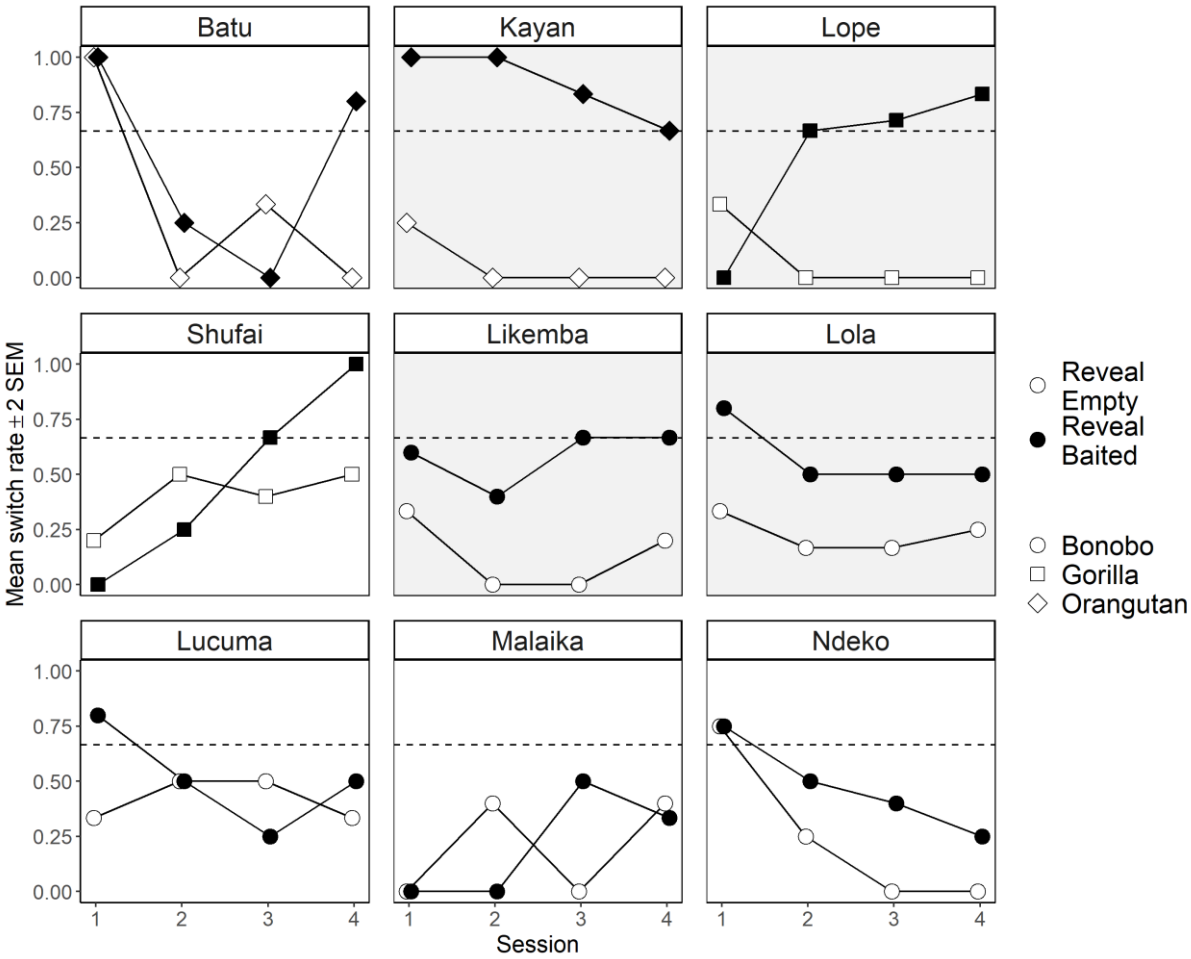


Figure 4.3.4: Individual switch rates for reveal empty and reveal baited trials across the 4 sessions of the two-choice condition for those individuals who switched at a rate significantly different from chance (binomial test, one-tailed, p = .05) in at least one trial type. Shaded boxes show those individuals who switched differentially based on condition (Fisher test, two-tailed, uncorrected p = .05)

The failure to detect a learning effect in experiment 1 is evidence against our data reflecting an associative strategy, however if we are to truly justify our conclusion that this is logical reasoning then we must ensure that subjects are able to apply it flexibly and selectively. In experiment 2, we will retest the 9 individuals who scored above chance in at least one condition but vary the baiting procedure to include trials where both food pieces are placed into one pair of cups. This was designed to account for a non-inferential strategy of *win-switch lose-stay*, whereby when a subject finds a food piece on their first guess they switch to the alternate set, but if they fail to find a piece, they should stay within the set. This will allow us to delineate those individuals who passed experiment 1 via reasoning logically, who will continue to perform well, from those who deployed this associative strategy, which will now be ineffective on 50% of trials. To assess the conditioning effects of experiment 1, we also tested a second group of chimpanzees who were experienced with search tasks but had not participated in experiment 1.

4.4 Experiment 2 – 4 cup control

Methods

Participants

The 9 individuals who scored higher than expected by chance in at least one condition of Experiment 1 progressed to Experiment 2(4 bonobos, 2 gorillas and 2 orangutans). To test whether the performance of this group was a conditioned response we also tested 9 chimpanzees at the Budongo Research Unit (BRU) in Edinburgh Zoo, these chimpanzees were research experienced but naïve to the experiment. Like the Twycross group, subjects are housed in species typical groups with access to indoor and outdoor spaces with vegetation, where testing takes place in a communal area accessible to all individuals during the testing period. The apparatus and procedure were the same but whole grapes were used as target items for the Edinburgh group, while the Twycross group continued to receive cubed sweet potato. To proceed to testing, subjects needed to score 8 out of 10 on the same memory pre-test task described in Experiment 1, 3 subjects failed to reach this criterion in two sessions. The final sample comprised 15 apes (7 female, mean age = 18.2), demographic details of all apes can be found in Table 4.4.1.

Table 4.4.1: Demographic details of participants in experiment 2.

Location	Species	ID	Sex	Age	Rearing
Edinburgh	Chimpanzee	Edith	Female	26	Parent
		Eva	Female	42	Parent
		Kilimi	Female	29	Parent
		Frek	Male	29	Parent
		Qafzeh	Male	30	Parent
		Velu	Male	8	Parent
Twycross	Bonobo	Likemba	Female	13	Unknown
		Lola	Female	3	Parent
		Malaika	Female	12	Parent
		Lucuma	Male	20	Parent
		Ndeko	Male	9	Parent
	Gorilla	Lope	Male	9	Parent
		Shufai	Male	5	Parent
	Orangutan	Kayan	Female	5	Parent
		Batu	Male	33	Parent

Apparatus

The apparatus was the same as experiment 1.

Procedure

Pre-test

The Edinburgh group performed the pre-test described in experiment 1, the Twycross group did no further pre-test and experiment 2 followed directly from experiment 1.

Test

In *standard* trials the procedure was identical to experiment 1. In *control* trials both food pieces were placed into one pair of cups, so subjects should choose both members of the baited pair. In control trials, the experimenter first showed the subject that all cups were empty and then baited the first pair as before, this time instead of placing the occluder in front of the second pair of cups, they placed it back down in front of the same pair and baited it again, so both cups within the pair contained a food item. This means that the un-baited pair were never covered and never touched by the experimenter. Each 12-trial session contained a block of six one-choice trials and a block of six two-choice trials, the order of which was counterbalanced between session. Within each 6-trial block there were four control trials and two standard trials, meaning that over the three sessions subjects received 12 control trials and 6 standard trials for each condition, for the one-choice condition these were split evenly between reveal-empty trials, where the experimenter selected from the unbaited pair, and reveal-baited trials where they removed a baited cup. Inter coder reliability for 15% of trials was perfect (Cohen's Kappa = 1).

Data coding and analysis

As in Experiment 1, analysis is based on the rate of switching between pairs, as the removed cup was not replaced, chance was set at 66%. For analysis of two-choice control trials, error types were coded as follows: *Correct both* - the subject correctly chose both cups from the baited pair; *Incorrect 1st choice* - the subject chose the un-baited pair on their first guess; *Incorrect 2nd choice* – the subject chose correctly on the first choice but switched to the alternate pair on their second choice; *Incorrect both* – the subject chose both cups from the unbaited pair.

Results and Discussion

Figure 4.4.1 shows how switch varied rates by condition in experiment 2. To test for differences in the likelihood of switching pairs, we fitted a mixed effects model with logit link function, using the predictors trial type (reveal empty/reveal baited), variant (one-choice/two-choice), condition (standard/control) and location (Edinburgh/Twycross) along with all interactions and the random effect of ID. The random effect did not improve the fit of the model over a GLM

with only the fixed effects (χ^2 = 1.364, df =1, p = .243), so we continued with the GLM. Similarly, we found no main effect of variant (one-choice/ two-choice) or support for any interactions containing it, so we removed it, which did not influence the model's fit (χ^2 = 10.40, df = 8, p = .238). The resulting model fit the data better than a null model without the predictors (χ^2 = 82.51, df = 15, p < .001). As the baiting procedure was different between control and standard trials, the correct response to each trial type (reveal empty or reveal baited) is reversed between conditions[22], so the interaction term is more informative than trial type alone.

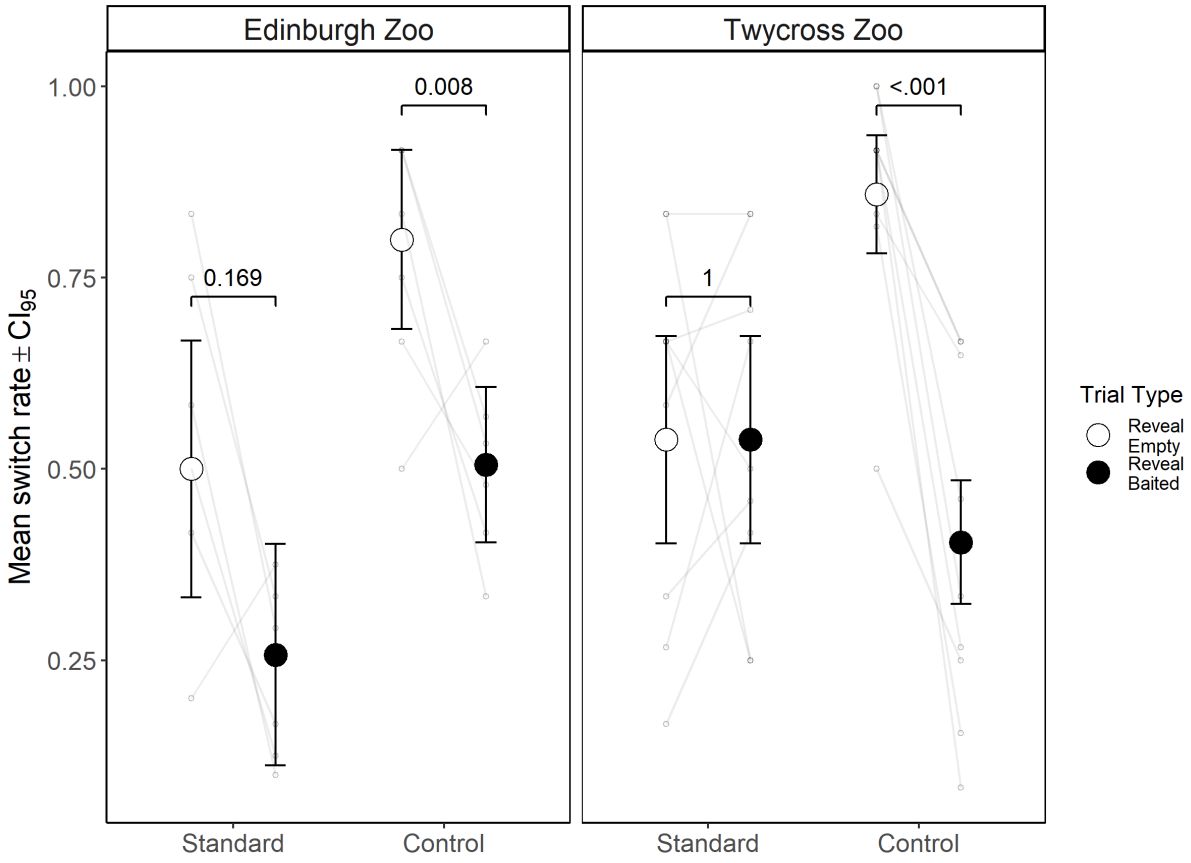


Figure 4.4.1: Estimated marginal means for a model to predict switching in Experiment 2 by trial type and group. Small circles show switch rates at the individual level. The variant term (one-choice vs two-choice) was not informative, so it was removed from the model (χ2 = 10.40, df = 8, p = .238).Pairwise comparisons can be found in Appendix 1, Table 8.3.1

Figure 3 shows the estimated marginal means from the resulting model by group, condition and revealed cup contents. We found that the effect of revealed cup contents varied by condition (χ^2 = 12.77, p < .001), which shows that subjects were switching differentially based on the baiting

[22] For example, in the standard condition there is one piece in each pair, so the correct response to reveal empty trials is to stay within the set as the revealed cup's partner certainly contains a piece, but in control trials subjects should switch because both cups in the alternate set were baited.

procedure and not simply responding based on a decision rule. However, while pairwise contrasts show that both groups switched differentially in the control condition (Edinburgh, β =1.365, CI_{95}(0.529,2.201), p = .008; Twycross, β =2.195, CI_{95}(1,474,2.915), p < .001) neither group did so in the standard condition (Edinburgh, β =1.06, CI_{95}(0.048, -2.07), p = .169; Twycross, β =0, CI_{95}(-0.771, 0.771), p = 1), This suggests that the presence of the control trials made the standard trials more challenging for the Twycross group, who switched differentially in Experiment 1[23] (paired t-test, $t_8 = 4.16$, p = .003).

Both groups significantly altered their choice rates between conditions in reveal empty trials (Edinburgh, β = 1.386, CI_{95}(0.394, 2.379) p = .032; Twycross, β = -0.542, CI_{95}(0.814, 2.492), p = .001) but not in reveal baited trials (Edinburgh, β =1.082, CI_{95}(0.222,1.943), p = .065; Twycross, β =-0.542, CI_{95}(-1.183,0.099), p = .346). Additionally, the 3-way interaction was significant ($\chi^2 = 4.85$, p = .028), showing that while the responses of the two groups to each condition are comparable, the magnitudes differ.

Particularly interesting is the performance of the Edinburgh group in standard trials, who switched approximately 50% of the time in reveal empty trials, which reproduces the performance of the Ngamba Island chimpanzees tested by Engelmann and colleagues (2022); but in reveal baited trials, subjects correctly switched sets on just ~25% of trials, compared to ~85% in the Ngamba group, suggesting that the location-based argument which the authors propose to explain their data is not sufficient to describe ours.

*Table 4.4.2: Coefficients from the model to predict switching sets in experiment 2. (switch~trial_type*condition*location)*

	Estimate	CI (2.5%)	CI (97.5 %)	p-value
(Intercept)	0.154	0.699	-0.391	0.579
(Trial type)Reveal Baited	0.000	0.771	-0.771	1
(Condition)Control	1.653	2.492	0.814	< .001
(Location)Edinburgh	-0.154	0.711	-1.020	0.727
(Trial type)Reveal Baited:(Condition)Control	-2.195	-1.139	-3.250	< .001
(Trial type)Reveal Baited:(Location)Edinburgh	-1.061	0.212	-2.334	0.102
(Condition)Control:(Location)Edinburgh	-0.266	1.033	-1.566	0.688
(Trial type)Reveal Baited:(Condition)Control: (Location)Edinburgh	1.891	3.576	0.206	0.028

[23] This comparison is based on individual rate is averaged across one-choice and two-choice trials of experiment 1, for the 9 individuals who participated in both experiments.

The data from both conditions suggest that the presence of control trials, where the correct response was reversed, made the standard trials more difficult. Crucially, this experiment has not found evidence that subjects are able to flexibly apply logical rules to solve variants of the same task, which suggests that logically reasoning via the disjunctive syllogism in a 4-cup 2-item task is outside of the capacity of great apes. This leaves us with the possibility that to pass the original two-choice variant of the task, the Twycross apes were using an associative rule, such as *win-switch lose-stay*, whereby if they are unsuccessful on their first guess they stay within the pair, but if they are successful, they switch to the alternate pair.

Analysis of errors in two-choice control trials.

Incorrect deployment of the *win-switch lose-stay* rule in two-choice control trials would result in the subject being either correct on their first guess but incorrect on their second or, if they were disregarding the baiting, being incorrect on both guesses. Figure 4.4.2 shows the distribution of errors in two-choice control trials. While the Edinburgh group made *win-switch lose-stay* errors more frequently (54.2% vs 37.0%), the overall distribution of error types did not vary between groups ($\chi^2 = 6.01$, df = 3, p = .111), which suggests that, while the error rate is surprisingly high, the Twycross group have not been conditioned into this response pattern.

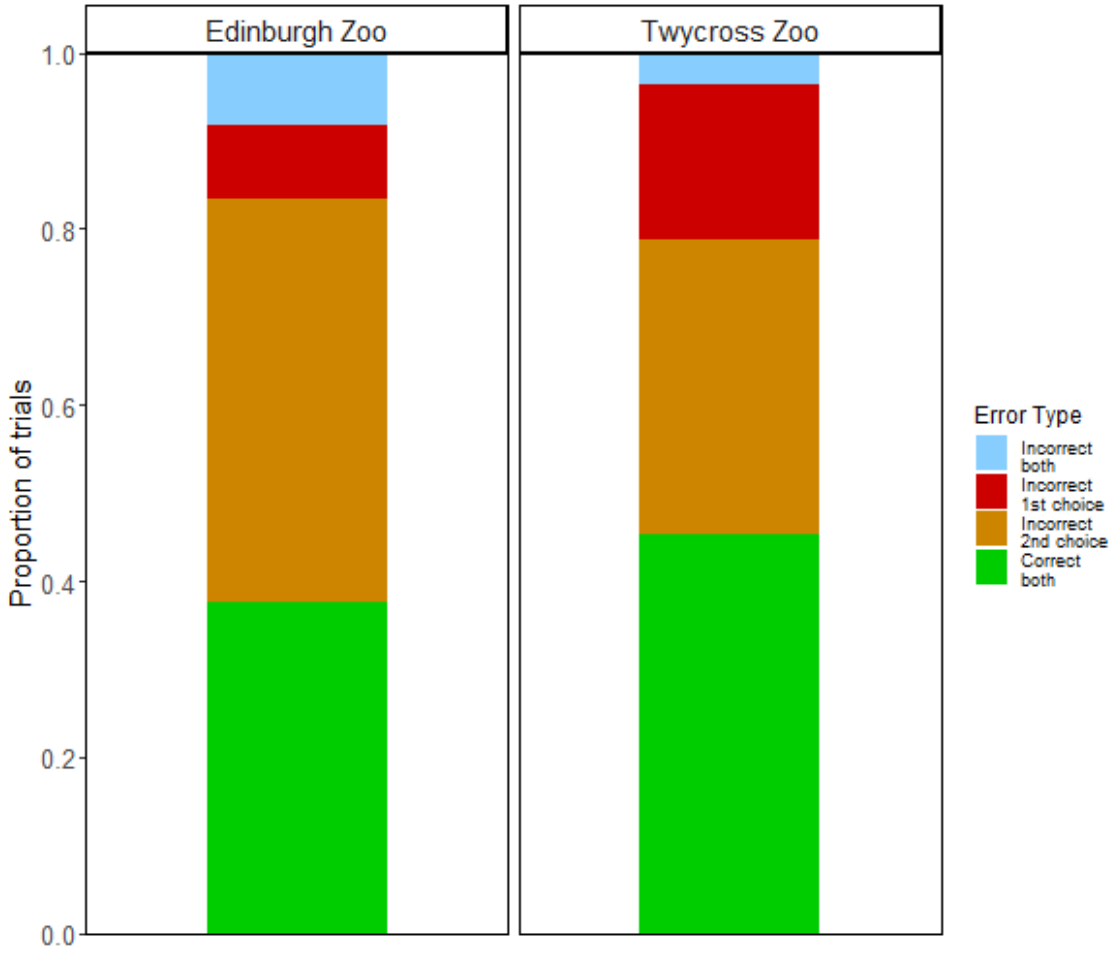


Figure 4.4.2: Distribution of error types in two-choice control trials of experiment 2 as a function of group.

Figure 4.4.3 shows the error distribution at an individual level in two-choice control trials. We see that Lope and Kayan, who had both passed the two-choice condition in Experiment 1, make errors on control trials indicative of having used a *win-switch lose-stay* strategy. However, while these errors mean that we must treat the results of Experiment 1 with caution, they are not necessarily terminal for a rich interpretation. These tasks involve a high number of trials, each of which require a large amount of cognitive effort, but in return the reward for each is relatively small. Therefor this paradigm is a viable candidate for the use of a heuristic, a simplified decision rule that serves to offer an approximately accurate answer in exchange for a drastically reduced cognitive load (Rieskamp & Hoffrage, 1999; Shah & Oppenheimer, 2008). While it is challenging to test for the formation of a heuristic, as it would manifest either as associative learning or in continuing to answer correctly, we can see it being superseded. In Kayan's first control session she made the *win-switch* error on 4/4 control trials, on the second session it was down to 2/4 trials and on the final session she no longer made the error.

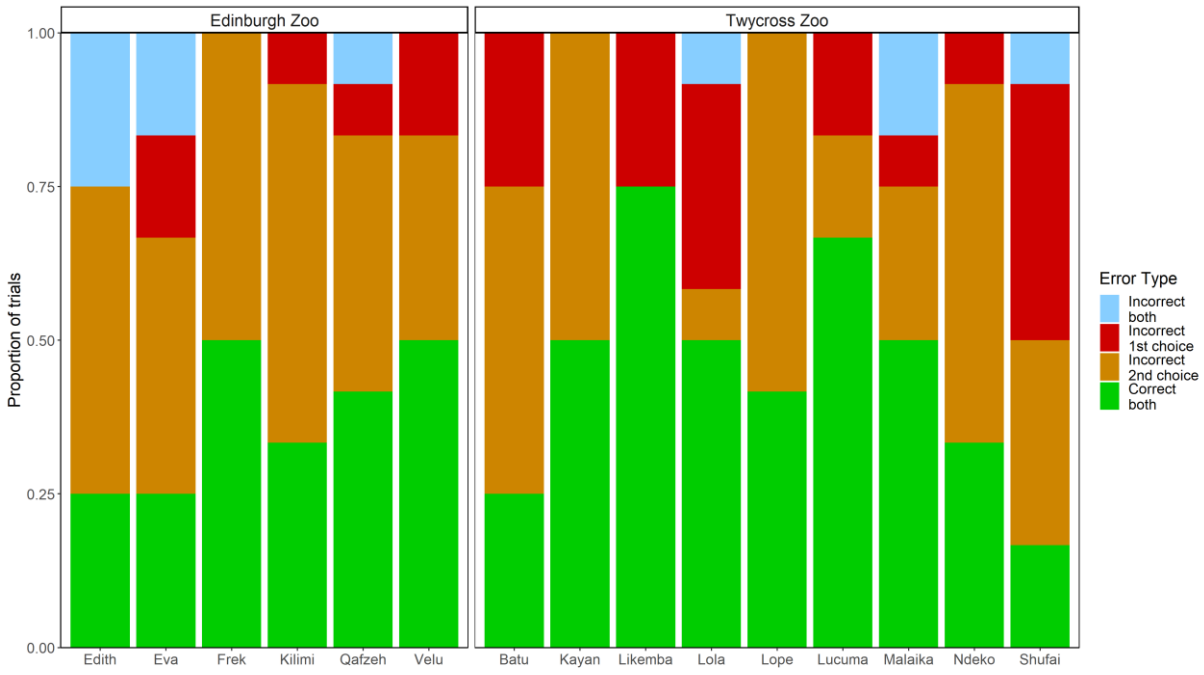


Figure 4.4.3:Individual level error distribution in two-choice control trials Individual level data can be found in Appendix 1 Table 8.3.2.

Previous literature has shown that primates also resort to heuristics under tasks with a high cognitive load (Broihanne et al., 2019), so it is plausible that, over the course of the 8 sessions, a cognitive strategy was replaced by a heuristic one, which the apes incorrectly applied in control trials. Crucially, because the baiting procedure was consistent throughout the first experiment, failing to attend closely to it was not penalised, possibly decreasing either the cognitive or attentional load for available inferential processes. This finds further support in the Twycross group's results in standard trials because, even though despite these standard trials being were identical to the that those of experiment 1, the additional cognitive demands placed on the subject by the mixture of trial types and conditions in rapid succession may have contributed to the decreased performance.

Nevertheless, when taken together, the results of Experiment 2 suggest that subjects were not solving Experiment 1 *logically*. Under a standard logic account (Rips, 2001) we apply logic sparingly to evaluate focussed parts of our model, so in a standard trial an agent can simply address the pair which is acted on in trials and would continue to perform well despite the addition of the control trials. Similarly, in control trials, the subject could eliminate two locations prior to the first cup being revealed and even a policy of simply avoiding the empty cup would be sufficient to pass the task.

4.5 General discussion

Across two experiments we have investigated whether great apes are able to reason logically in line with an understanding of the disjunctive syllogism. Firstly, we found that, as a group, subjects adaptively switch between pairs depending on the contents of the revealed cup in both the one- and the two-choice paradigm. Secondly, we have shown that great apes perform better in trials that test the inclusive than the exclusive disjunction (compared against chance responding). At the individual level, we have shown that one subject was able to reason via both the inclusive and the exclusive disjunction in a 4-cup task, demonstrating that language is not required reason logically under this paradigm. Using a second experiment involving those subjects who passed the first experiment along with a group of naïve subjects, we showed that their responses are not a learned heuristic, but they were likely based on paying attention to both the baiting and the outcome.

However, the addition of control trials resulted in a decrease in performance compared to experiment 1, meaning that apes were not able to flexibly apply the inference needed to solve the standard trials in experiment 2.

Both of these groups have previously passed a 2-cup disjunctive syllogism task (Chapter 3), which suggests that performing the inference operation is not a limiting factor in apes' performance. When we combine this with the observation that some subjects are able to pass this task without being able to flexibly apply the same logical operation, it would suggest that the 4-cup task is not a suitable test for reasoning ability in great apes.

It is possible that the working memory constraints required to track 2 food items across 4-possible locations could be constraining performance. For example, an alternative model in which the subject only attends to one food piece would also make the same predictions as both the minimal- and location-based models. In reveal empty trials subjects would simply select from the set containing the piece which they had been attending to (their target set), resulting in a 50% switch rate; and in reveal baited trials subjects would either select from their target set or, if that item had been removed, they would switch to the alternate pair. Alternatively, if they had completely disregarded the other piece and were to choose randomly between the 3 cups, we

would see a resultant switch rate of 83% in reveal baited trials[24], which closely approximates the rate reported by Engelmann et al. (2022).

While working memory constraints cannot be the explanation for infants, who can track small sets (Feigenson & Carey, 2003) and keep the resulting representations separate (Rosenberg & Feigenson, 2013) but still fail the 3-cup (Leahy et al., 2022) and 4-cup tasks (Mody & Carey, 2016); however that an explanation cannot also be applied to children does not rule out its utility, as divergence between apes and 3-year-old children on the 4-cup task already indicates that different elements of the task are the limiting factor in performance for these groups. As suggested by Krupenye and Call (2019) when discussing why great apes had failed early, food-based theory of mind tasks, the inhibitory and attentional demands which food places on apes may be masking cognitive abilities[25]. While no explanation is sufficient in isolation, it may be that the cognitive demands placed on the subject by multiple elements of the task are contributing to great apes' failure.

From a human perspective, the argumentative theory of human reasoning (Mercier and Sperber 2010) suggests that reasoning originally evolved for a communicative purpose, as an interactive strategy to critique the merits and flaws of different plans and strategies. Subsequently, individual reasoning emerged as an introspective simulation of inter-personal reasoning. Thus, the argumentative theory places language as an integral aspect in the emergence of reasoning, but crucially, individual reasoning acts on intuitions formed from environmental regularities, which are available without the power of language. These are the building blocks on which inference takes place, so either the recognition of these environmental regularities, or the ability abstract them out of the specific scenario in which they occur could be the basis for the variation which we have observed between subjects.

The language centric definition is reminiscent of the distinction between inductive and deductive inference, both of which share the end-goal of using held knowledge to derive new knowledge, where they differ from one another is the method by which they reach the new knowledge. Deductive reasoning uses strict rules or axioms, statements which are known to be universally true, to make new prepositions from initial premises; induction, by contrast, does not require formal rules and instead works via drawing logical conclusions based upon previous

---

[24] In half of the trials the experimenter would act on the non-target set, to which the subject would respond by searching for their target piece, resulting in a 100% swich rate; in the remaining half of the trials the experimenter would remove the subject's target piece, if the subject were to then randomly choose from the remaining cups they would switch on 66% of trials, for a combined switch rate of 83%.
[25] Failure to inhibit searching for a randomly specified piece would be reflected in a 50% switch rate on reveal empty trials and 100% of reveal baited trials for same reasons as described above.

observations (Henderson, 2020). The characteristic difference being that deductive inference relies on absolute truths, while inductive inference relies upon probabilities.

While non-human animals may be capable of tracking these environmental regularities, having intuitions, and making rational choices based on them; these conclusions are likely still probabilistic, as deductive reasoning is an entirely language-based concept. While a notable peril of experimental primate cognition research is that subjects are exposed to various paradigms over time thereby endangering carryover between tasks, this may only place them on a par with the exposure that a developing infant receives, thus allowing them the experience to make the abstract rules ubiquitous in human development.

Grigoroglou and Ganea (2022) point to the polysemy of many of the modal verbs and argue that children learn the simpler non-epistemic uses of the word first, which they then use to scaffold the modal concept later. They authors note that children do not start to use the semantic meaning of the word in an adult sense until the age of 7, it is plausible that the young children tested by Mody and Carey (2016) and by Gautam, Redshaw and Suddendorf (2021), are also responding simply with intuitions. However, the question remains as to why reasoning via the disjunctive syllogism in a 4-cup task is near-ubiquitous in children by the age of 5, but only present in a handful of apes tested. The possibility which always exists in comparative research is that there are task constraints which are limiting performance, and redesigned paradigms or testing behaviour indirectly may provide more answers. To this end, investigating deductive inference using a visual search paradigm may not be possible and better answers could come from novel physiological measurements that can characterise violations of expectation.

Conclusion

Here we have shown that, under both the one-choice and the two-choice 4-cup paradigm, great apes switch adaptively in line with reasoning via the disjunctive syllogism. This adaptive switching is driven largely by increased performance in the inclusive disjunction, in which subjects switched at a rate lower than chance, which may suggest that they are failing to represent the exclusive nature of the disjunction. At an individual level, two individuals switched adaptively based on the contents of the revealed cup and their results cannot be explained by the minimal model. Thus, we have shown that reasoning via both variants of the disjunctive syllogism in a 4-cup task is not unique to humans. However, when we include associative control trials, which were absent from previous studies, we find that performance breaks down, suggesting that apes are not able to flexibly apply the underlying logical operation. When we consider this finding alongside the evidence that these apes can reliably

solve a 2-cup disjunctive syllogism task, we conclude that the 4-cup task may not be an effective measure of logical reasoning in non-human primates.

5. Counterfactual curiosity: motivation to know what was once possible.

5.1 Abstract

When non-human primates make a choice, do they consider what could have happened if they chose differently? We used a metacognition paradigm to test whether after finding the outcome of an uncertain choice, chimpanzees would move to check the contents of the unchosen option. We found that chimpanzees preferentially checked the location which they could have chosen but didn't, over one which was never available to them (Experiment 1), and that this couldn't be explained by a reductive, proximity-based account (Experiment 2). In Experiment 3 we manipulated whether subjects had agency over the outcome and found that, while chimpanzees only showed a bias towards the available location in the agency condition, having agency did not change the rate with which they checked the counterfactual. Taken together, these data suggest that non-human primates are specifically curious about counterfactuals and not simply motivated to resolve uncertainty, but that agency does not hold the same significance as in humans.

5.2 Introduction

For better or worse, humans expend a lot of energy thinking about things that could have been. These alternative versions of reality are known as counterfactuals. While adults and children alike relish in creating rich fantasy worlds (Dubourg & Baumard, 2022), the counterfactual scenarios on which we ruminate are often egocentric and grounded in reality. As they demand so much of our attention, it is natural to ask whether these ruminations are uniquely human. While counterfactual simulation is understood to be an essential aspect of how humans learn about the world through our actions (Epstude & Roese, 2008; Seligman et al., 2013; Weisberg & Gopnik, 2013) multiple authors have suggested that non-human animals are 'stuck in the present' and unable to reason about either the past or the future (Hoerl and McCormack, 2019; Redshaw and Suddendorf, 2020; Suddendorf and Corballis, 2007; c.f. Corballis, 2019).

Mature counterfactual reasoning requires that an individual ignores the factual state of the world that they know to be true to simulate it in alternate state (Byrne, 2016), an ability which is thought to emerge around six years of age in children (Beck, 2016; McCormack et al., 2018; Nyhout et al., 2019). However, others have suggested it to be as early as 4 years of age (Nyhout

& Ganea, 2019) or as late as 12 years (Rafetseder et al., 2013). While counterfactual reasoning requires simulation, counterfactual curiosity does not, and is instead defined as simply seeking information about alternative past possibilities. It is thought to be less cognitively demanding and has been proposed to emerge earlier in human development (FitzGibbon et al., 2019).

In an card matching task that allowed pre-school children the option to use 'x-ray glasses' to find out what was under the card that they didn't choose, 75% of children used the glasses even in trials where there was no option to change their answer (FitzGibbon et al., 2019). Similarly adults, despite being made aware of the absence of instrumental value, will 'pay' for counterfactual information using time, effort and money (FitzGibbon et al., 2021). Like the human participants, rhesus macaques *(Macaca mulatta*) will pay a nominal cost for counterfactual information in the absence of future utility (Wang & Hayden, 2019). When selecting between two gambles, the macaques preferentially selected the gamble that provided them with information about their unchosen option even if the odds were worse, and their willingness to pay varied as a function of information availability. The negative cost-benefit balance of counterfactual curiosity makes its persistence somewhat enigmatic. Two arguments that have been made to explain its existence, mechanistic and functional (Fitzgibbon & Murayama, 2022).

The mechanistic argument is that curiosity functions to decrease uncertainty through active sampling. According to this idea, even the simplest organisms are intrinsically motivated to decrease uncertainty about the world (Friston, 2010; Gottlieb & Oudeyer, 2018; Iigaya et al., 2016). This mechanistic argument is sufficient to explain the behaviour of the macaques described above (Wang & Hayden, 2019). Stigler (1961) argued that the pursuit of information was strictly valuable to the extent to which it permits us to make better decisions, yet the intrinsic pursuit of information is documented in a number of species (Bromberg-Martin & Hikosaka, 2009; Eliaz & Schotter, 2007; Vasconcelos et al., 2015). As such, decreasing uncertainty must have some value, else it would not have evolved. The crucial difference between uncertainty resolution and counterfactual curiosity is that information search of a counterfactually curious agent is strategically directed towards alternative past possibilities, rather than indiscriminately.

The functional argument is that counterfactual information often does contain information with instrumental value, so throughout development we learn an association between the two. From this perspective curiosity is not just intrinsically motivated (Kidd & Hayden, 2015; Loewenstein, 1994) but driven by its positive consequences. It is highly likely that both

elements play a role in human counterfactual curiosity so simply demonstrating counterfactual curiosity does not distinguish between them.

As failure provides more opportunity for learning, the functional argument predicts an increase in information search after negative outcomes. A second card matching experiment by Fitzgibbon et al. (2019) manipulated agency by controlling whether subjects chose from one or two cards at their first choice. The authors found that the pre-school subjects were most likely to bias their search towards the unchosen option after negative results which they had agency over[26], rather than positive or chance outcomes.

Bault et al. (2016) gave adult subjects a choice between two spinners and showed that after a negative outcome subjects focussed more, as measured by looking time, on the realised outcome of the unchosen gamble, rather than the unrealised outcome of the chosen one. This suggests, and emotional ratings confirmed, that agency played a particular role in their feelings of regret. Guerini et al. (2020) replicated the two spinner experiment with children aged 3-10 but manipulated whether the subject or a computer selected which spinner they would play with. As the aim of the task was to investigate the counterfactual emotion regret rather than counterfactual curiosity, the authors did not measure looking time. However, emotional ratings showed that agency only starts to impact children's experience of regret after the age of 6, considerably later than the pre-school children tested in the card matching game.

In the current study we tested chimpanzees' counterfactual curiosity by giving them a choice from 2 of 3 differentially baited locations. This means that knowing what they received didn't inform them what they passed over. After the trial was over, we uncovered the ends of the hiding locations and recorded whether the chimpanzees moved to peek into either of the 'unchosen' or the 'unavailable' locations, an information seeking behaviour that was originally used to investigate metacognition (Call & Carpenter, 2001). If subjects are simply resolving uncertainty, then we should see no difference in the proportion of checks directed at the unchosen and unavailable tubes, but a bias towards the unchosen tube would show that they are specifically curious about the counterfactual. Moreover, analysing whether they use this information in the next trial will inform us about whether they are showing curiosity, or simply searching for information. In the second experiment we used an information seeking paradigm to test whether subjects preferentially search tubes which are closest to them. Finally, we manipulated whether the subjects had agency over which tube they received in order to

[26] In analysis of the no agency condition one of two unavailable cards was selected to be the unchosen card, which served to act as a control.

investigate whether, like humans, they would devote greater attention to the counterfactual option after negative outcomes which they had agency over.

5.3 Experiment 1

Methods

Subjects

We tested 11 chimpanzees (*Pan troglodytes*) living at the Budongo Research Unit (BRU) in Edinburgh Zoo. The subjects live in a natural group with indoor and outdoor access, they are provided regular produce feedings and additional enrichment. Testing takes place in a communal area with water available ad libitum and subjects are never separated from the group for research purposes. Individuals are experienced in non-invasive cognitive testing including search paradigms and are familiar with tubes but have not engaged with comparable research. Three individuals failed the pre-test, so we entered the test phase with 8 participants (3 female), ranging in age between 3 and 42 (mean = 25.1 years) and all mother-reared (Table 5.3.1).

Table 5.3.1: Demographic details of participants

ID	Age	Sex	Rearing History
Eva	42	F	Parent
Frek	29	M	Parent
Kilimi	30	F	Parent
Masindi	3	F	Parent
Paul	29	M	Parent
Qafzeh	30	M	Parent
Rene	30	M	Parent
Velu	8	M	Parent

Apparatus

We presented the subjects with 3 tubes (L = 400mm, Ø = 40mm) spaced equidistantly and oriented in parallel on a sliding table (630mm x 300mm) between the subject and the experimenter such that the subjects could peek into the tubes' open ends. We covered the ends of the tubes to block visual access using small freestanding barriers (100mm x 50mm), and during baiting covered the whole table with a large box. We used a large apple piece ($1/12^{th}$ of an apple) and a small piece ($1/36^{th}$ of an apple). The small piece represents the standard size for cognitive tasks at the BRU, so the large piece was a windfall.

Procedure.

Pre-test:

In the pre-test phase, chimpanzees were presented with the three tubes on the table with their ends uncovered. The experimenter (E) then placed the box over the whole table and placed a large reward, and a small reward onto the box above the left two tubes. The E then lifted the large piece to show it to the subject and placed it into the end (closest to the experimenter) of one of the three tubes, visiting all tubes in a left-right direction, and repeated this procedure with the small piece. The E then removed the box and slid the table to the subject. This task essentially amounted to an information seeking paradigm. To pass the pre-test subjects had to select the large piece on 8 out of 10 trials within 2 sessions. One juvenile individual, Masindi, received an additional pre-test session in which the tubes were instead baited with a small apple piece and a blueberry, a lower value food item, to counter her initial response of selecting the first baited tube which she encountered. After which she passed the standard pre-test.

Test:

The baiting procedure of the test phase was the same as the pre-test with the exception that the ends of the tubes (closest to the subject) were covered. After baiting, the E lifted the box to reveal the tubes, paused and dragged one of the tubes to the back of the board. The E then slid forward the table for the subject to make their choice between the two remaining tubes, which they did so via one of three choice holes at the bottom of the plexi-glass panel. The two tubes available for choice are referred to as the *chosen* and *unchosen* tubes, and the tube pulled to the back of the board is referred to as the *unavailable* tube.

After the subject made their choice, the E lifted their chosen tube, turned it, and shook it so that its contents fell into their hand. If it contained an apple piece the experimenter gave it to the subject before placing the tube back onto the table in its original position. The E waited 3 seconds and then removed the barriers covering the end of the other two tubes (unchosen and unavailable) then waited 5 seconds to allow the subjects to look into the tubes, and then reset the trial. Figure 5.3.1 shows this procedure. If subjects had checked both tubes before the 5 seconds, the experimenter ended the interval early. If the subject was displaced by another individual the trial was repeated but if they left the testing window of their own accord the trial was coded as standard.

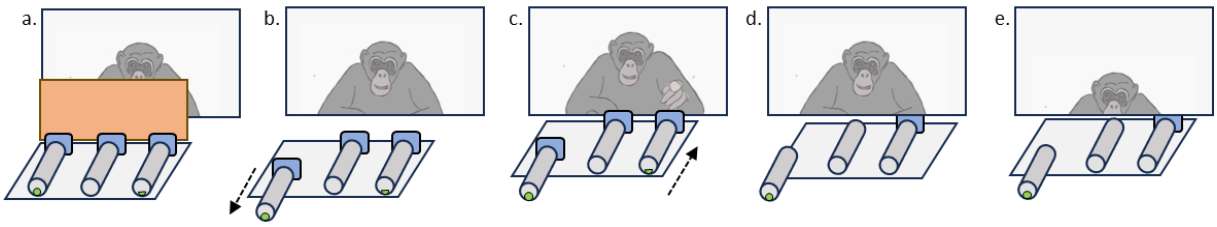


Figure 5.3.1: Procedure for test trials (left to right). E baits tubes behind a barrier (a), E removes barrier and makes one tube unvailable by sliding it backwards (b), E slides table to subject to make their choice (c), E pauses 3 seconds before removing covers from unchosen and unavailable tubes (d), E pauses 5 seconds to allow subject to (optionally) check uncovered tubes (e).

Each subjects received 3 sessions of 8 trials for a total of 24 trials. The contents of each tube and the identity of the tube that was made unavailable were combined to make 12 possible orientations, each of which was repeated twice. The sequence of trials was pseudorandomised to approximately counterbalance the identity (1-3) and the contents of the unavailable tube within and between the 3 sessions. Subjects were then randomly assigned which 8-trial block (1-3) they would start on.

Data coding and analysis

Each trial was coded from video recordings taken from a camera placed next to the experimenter at 80cm above the floor and in line with the right edge of the table. Subjects were considered to have checked a tube if they moved their head below a mark placed at 10cm above the table-top and in-front of the relevant tube. This was the height at which an animal keeper could see an apple piece at the far end of the tube while 30cm from the plexi-glass panel. A second experimenter scored 25% of the trials. Inter-observer reliability was excellent (kappa = .973). All statistics are paired unless otherwise stated.

Results and Discussion

All individuals checked at least one tube on at least one trial. The left pane of Figure 5.3.2 shows the proportion of trials in which subjects looked into the unchosen and unavailable tube within the 5 second time frame. Subjects looked into the unchosen tube significantly more often than the unavailable tube (t-test, $t_7 = 3.47$, $p = 0.013$), which suggests that they were curious about possible outcomes rather than simply the contents of the tubes.

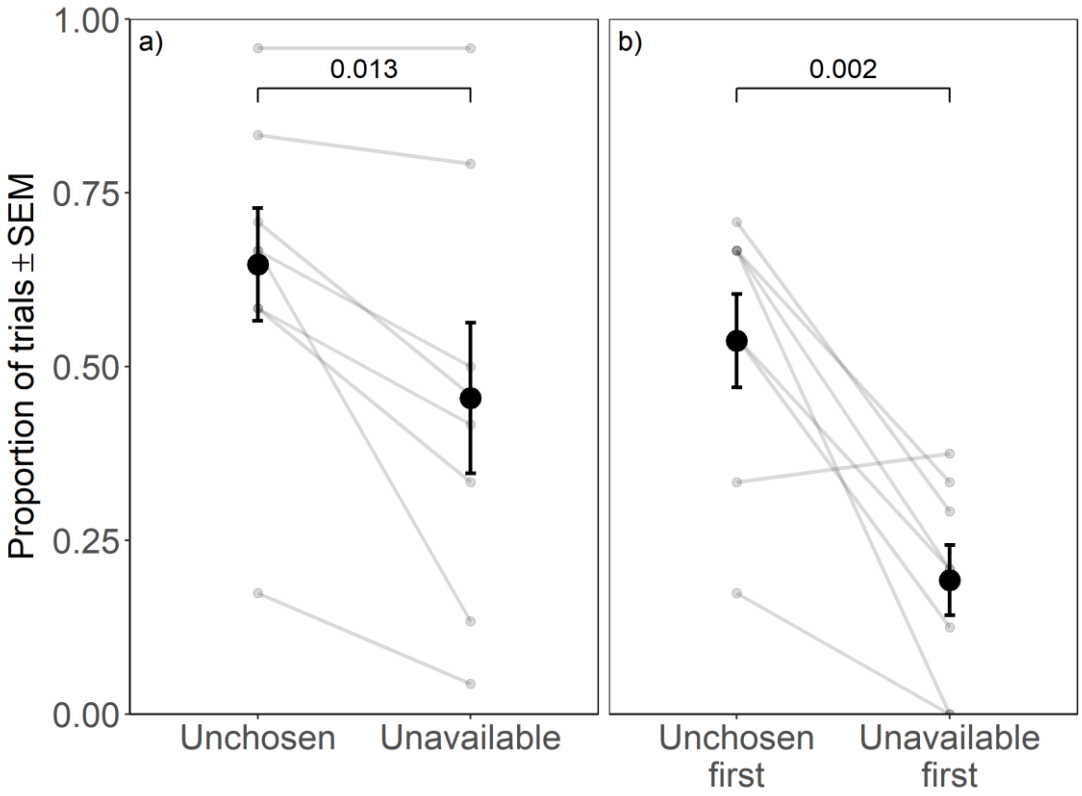


Figure 5.3.2: Mean proportion of trials in which subjects looked into the unchosen and unavailable tubes (left). Mean proportion of trials in which subjects directed their first look at the unchosen and unavailable tubes (right).

As the cost of checking the tubes was low, on 37% of trials the subject checked both tubes. The right pane of Figure 5.3.2 shows the proportion of first checks[27] directed at each tube, once again demonstrating a significant attentional bias towards the unchosen tube (t-test, $t_7 = 4.66$, $p = 0.002$). This is consistent with a literature which demonstrates that great apes are discriminate with their metacognitive checks, checking more frequently when the stakes are higher or the interval longer (Call, 2010), rechecking when they receive conflicting evidence (O'Madagain et al., 2022) and not continuing to search once they have found the reward (Engelmann et al., 2021).

To test the hypothesis that this was information seeking instead of curiosity, we analysed whether subjects chose the tube which had contained the windfall piece on the last trial. Subjects chose that location on 56.5% of trials where it was available, and there was no difference in proportion of guesses directed towards that location between trials where subjects had or had not checked that tube in the last trial ($\chi^2 = 0.045$, df = 1, $p = .831$). Therefore, we can

[27] If a subject checked both the unchosen and chosen tube, only their first look was coded.

reject that subjects were seeking information that they used to inform their choices in the next trial.

Table 5.3.2: Individual rates of checking each tube in Experiment 1.

	Check Neither	Check Unchosen	Check Unavailable	Check Both	Unchosen First	Unavailable First
Eva	0	0.833	0.792	0.625	0.708	0.292
Frek	0.25	0.583	0.333	0.167	0.542	0.208
Kilimi	0.125	0.708	0.458	0.292	0.667	0.208
Masindi	0.292	0.667	0.5	0.458	0.333	0.375
Paul	0.826	0.174	0.043	0.043	0.174	0
Qafzeh	0	0.958	0.958	0.917	0.667	0.333
Rene	0.333	0.667	0.133	0.133	0.667	0
Velu	0.333	0.583	0.417	0.333	0.542	0.125

Here we have demonstrated that chimpanzees are curious about counterfactual outcomes and preferentially direct their attention towards options which they could have selected, the nearest possible counterfactual. They do so after the trial has ended and don't use that information in the following trial, demonstrating that this is, in fact, curiosity rather than information seeking. Directing their search towards the unchosen rather than the unavailable tube lends support to the conclusion that subjects seek counterfactuals rather than simply resolving uncertainty, in which case we would have seen looks directed equally to both tubes. Moreover, engaging with the *nearest possible counterfactual*, that is, to hold all else constant apart from the counterfactual condition and its downstream consequences, is an element that differentiates mature counterfactual reasoning from basic conditional reasoning (Edgington, 2011; Rafetseder et al., 2010). Basic conditional reasoning involves the insertion of general regularities from outside the vignette, such as that removing dirty shoes is associated with clean floors remaining clean, and can lead young children to answer questions in a way that resembles counterfactual reasoning (Leahy et al., 2014). This is why some authors consider that full counterfactual reasoning develops at a later age (Rafetseder et al., 2013). While we are not equating counterfactual curiosity with counterfactual reasoning, conforming to its constraints is essential for proposing a continuity between the two.

However, an alternative explanation for our results is that subjects simply preferred to first check the tube that was physically closer to them (the unchosen one). We tested this hypothesis in Experiment 2 by repeating the baiting procedure of Experiment 1 except that we left one of

the available tubes (closest to the subject) and the unavailable tube uncovered. So, subjects faced a close covered tube, a close uncovered tube and a far uncovered tube. If subjects' tube inspections were based on proximity to the tube, we predicted that subjects would show a preference for the nearest tube.

5.4 Experiment 2

Methods

The same subjects took part in Experiment 2. The sequence of trials for each individual was identical to the sequence that they received in Experiment 1 and the identity of the covered tube (1-3) was yoked to their choice in the corresponding trial of Experiment 1. This meant that the layout of the apparatus before the choice in Experiment 2 was identical to that of post-decision inspection phase of Experiment 1.

The trial started with all 3 tubes uncovered and empty, the E then placed the large box over the top of the tubes and baited them as in Experiment 1. After baiting, the experimenter slid the unavailable tube back and placed a single cover over the subject's choice from Experiment 1. The E then removed the box and moved the table to its forward position, to allow the subject to (optionally) inspect the open tubes and then make their choice. Inter-coder reliability for 25% of trials was excellent (kappa = 0.936). Figure 5.4.1 shows this procedure diagrammatically.

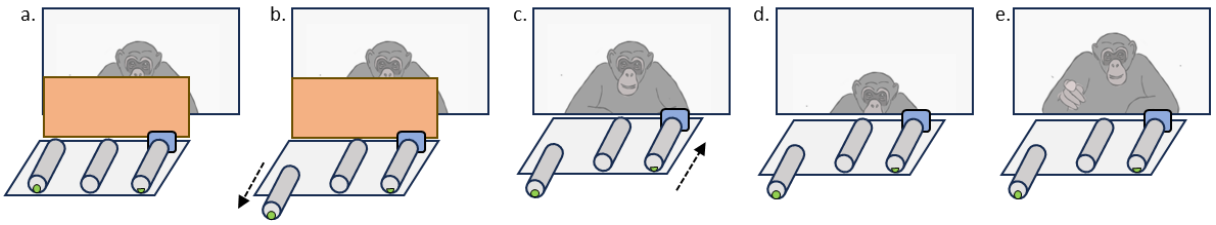


Figure 5.4.1: Procedure for experiment 2, (left to right). E baits tubes behind barrier and places cover over subject's choice from experiment 1 (a), E slides one tube backwards (b). E slides table to subject (c),(optional) subject checks uncovered tubes (d), subject chooses a tube and E gives them its contents (e).

Results and Discussion

We found that subjects checked both uncovered tubes equally (t-test, $t_7 = -0.794$, $p = .453$) and show no difference in which tube they checked first (t-test, $t_7 = -1.660$, $p = .141$). This means that we can reject this distance-based explanation for the effect found in Experiment 1. Individual level data can be found in Appendix 1 Table 8.4.1.

5.5 Experiment 3

In Experiment 1 we demonstrated that chimpanzees preferentially checked the option that they could have chosen rather than the one that was made unavailable to them by the E. In Experiment 3 we tested whether the act of making the choice between the two available tubes increased the likelihood of checking the unchosen tube and whether this was modulated by the contents of the chosen tube. We tested this idea by measuring whether checking was higher in a condition in which subjects chose one of the tubes compared to a condition in which the E chose for them. Secondly, to test for targeted information seeking, we altered the paradigm from experiment 1 to instead uncover the unchosen and chosen tubes, thus checking the tubes would now represent an informative and an uninformative search.

Methods

One individual did not participate in Experiment 3 leaving seven subjects in our sample (3 female, mean age: 24.4 years). The baiting procedure and target items for Experiment 3 were the same as Experiment 1. In the *agency* condition the choice procedure was the also same as Experiment 1 - the E slid one tube back and then slid the whole table forward for the subject to make their choice between the remaining tubes before giving the subject its contents. In the *no agency* condition, after sliding the unavailable tube backwards the experimenter paused before picking up one of the tubes and giving the subject its contents, then sliding the table to its forward position. After pausing 3 seconds for the subject to eat their piece, the experimenter then removed the covers from the chosen- and unchosen tubes. They then waited for interval of 5 seconds before resetting the trial, if the subject clearly checked both tubes, the experimenter reset the trial early.

Subjects again received 3 sessions of 8 trials, with each orientation (tube configuration and contents) being tested once with and without agency. The tube contents and identity of the unavailable tube were counterbalanced and pseudorandomised as in experiment 1. The location (positions 1-3) and contents (large, small, nothing) of the tube chosen by the E was counterbalanced between trials. Trials alternated between agency and non-agency and subjects were randomly assigned to start one or the other condition. As in experiment 1, subjects were also randomly assigned to a block order. Inter-coder reliability for 25% of trials was excellent (kappa = .908).

Results and Discussion

Figure 5.5.1 shows the rate of checking the chosen and unchosen tubes as a function of agency. Pairwise t-tests (with Bonferroni correction), show us that in the agency condition the subjects again show a bias towards checking the unchosen tube (t-test, $t_6 = 3.00$, $p = 0.0479$), however this bias disappears in the non-agency condition (t-test, $t_6 = 0.484$, $p = 1$). Since checking the unchosen and chosen tubes amounts to an informative vs an uninformative search, we conclude that subjects only engage in directed search in the agency condition. When analysing only first looks (Figure 5.5.2), we found no overall effect in the agency condition (t-test with Bonferroni correction, $t_6 = -2.42$, $p = 0.103$), nevertheless inspection of the individual level data would suggest that for most individuals, this relationship holds.

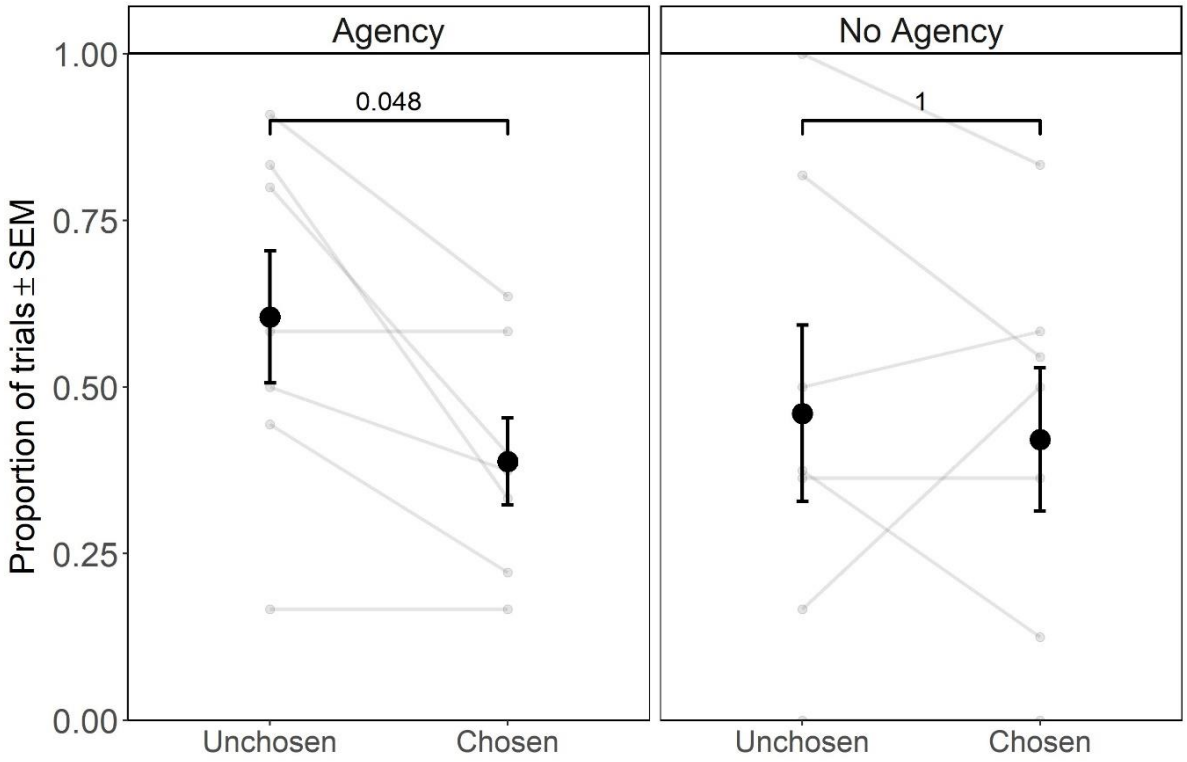


Figure 5.5.1: Mean proportion of trials in which subjects looked into the unchosen and unavailable tubes in experiment 3, as a function of agency, small points show individual level data. Annotations show the adjusted p-values from paired t-tests with Bonferroni corrections for multiple comparisons

Table 5.5.1: Individual rates of checking each tube in experiment 3, as a function of agency.

Condition	ID	Check Neither	Check Unchosen	Check Unavailable	Check Both	Unchosen First	Unavailable First
Agency	Eva	0.167	0.583	0.583	0.333	0.333	0.5
	Frek	0.167	0.833	0.333	0.333	0.667	0.167
	Kilimi	0.2	0.8	0.4	0.4	0.8	0
	Masindi	0.556	0.444	0.222	0.222	0.333	0.111
	Paul	0.75	0.167	0.167	0.083	0.167	0.083
	Qafzeh	0	0.909	0.636	0.727	0.636	0.364
	Velu	0.5	0.5	0.375	0.375	0.375	0.125
No agency	Eva	0.333	0.5	0.583	0.417	0.083	0.583
	Frek	0.091	0.818	0.545	0.455	0.636	0.273
	Kilimi	0.333	0.167	0.5	0	0.167	0.5
	Masindi	0.545	0.364	0.364	0.273	0.182	0.273
	Paul	1	0	0	0	0	0
	Qafzeh	0	1	0.833	0.833	0.75	0.25
	Velu	0.5	0.375	0.125	0	0.375	0.125

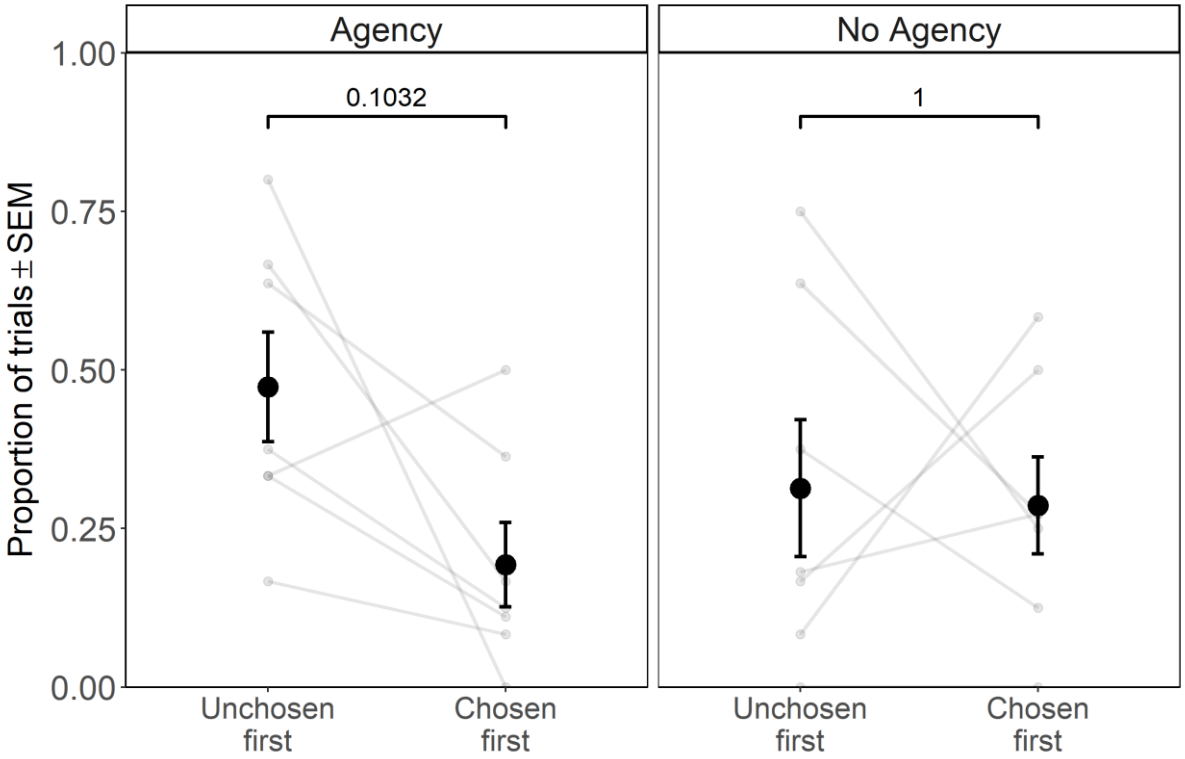


Figure 5.5.2:Mean proportion of trials in which subjects directed their first looks at the unchosen vs the chosen tube, as a function of agency, lines and small points show individual level data. Annotations show the adjusted p-values from paired t-tests with Bonferroni corrections for multiple comparisons.

The analysis so far has shown that having agency over the choice biases behaviour towards the informative search but has not answered whether agency impacts the frequency with which subjects search for information. To investigate how the binary outcome of checking the unchosen tube varied as a function of agency and outcome we fitted a GLMM (package *lme4*) with the fixed effects of condition (agency vs no agency) and outcome (as an ordinal factor, nothing < small < large), and the random effect of ID. Our model was an improvement over our null model containing only the random effect ($\chi^2 = 13.634$, $p = .003$) and a GLM containing only the fixed effects ($\chi^2 = 39.26$, $p < .001$). We found a main effect of outcome both linearly ($\beta = -0.998$, $CI_{95}(-1.828, -0.225)$, $p = .013$) and quadratically ($\beta = 0.764$, $CI_{95}(0.048, 1.525)$, $p = .039$), suggesting that subjects discriminated between nothing and something, but not between the two sizes of apple piece (Figure 5.5.3). We found no effect of agency ($\beta = -0.458$, $p = .287$) and the addition of the agency-outcome interaction did not improve the fit of the model ($\chi^2 = 0.138$, $df = 2$, $p = .934$). Thus, while agency biases chimpanzees' search towards the unchosen tube, it does not impact the absolute frequency with which they check it.

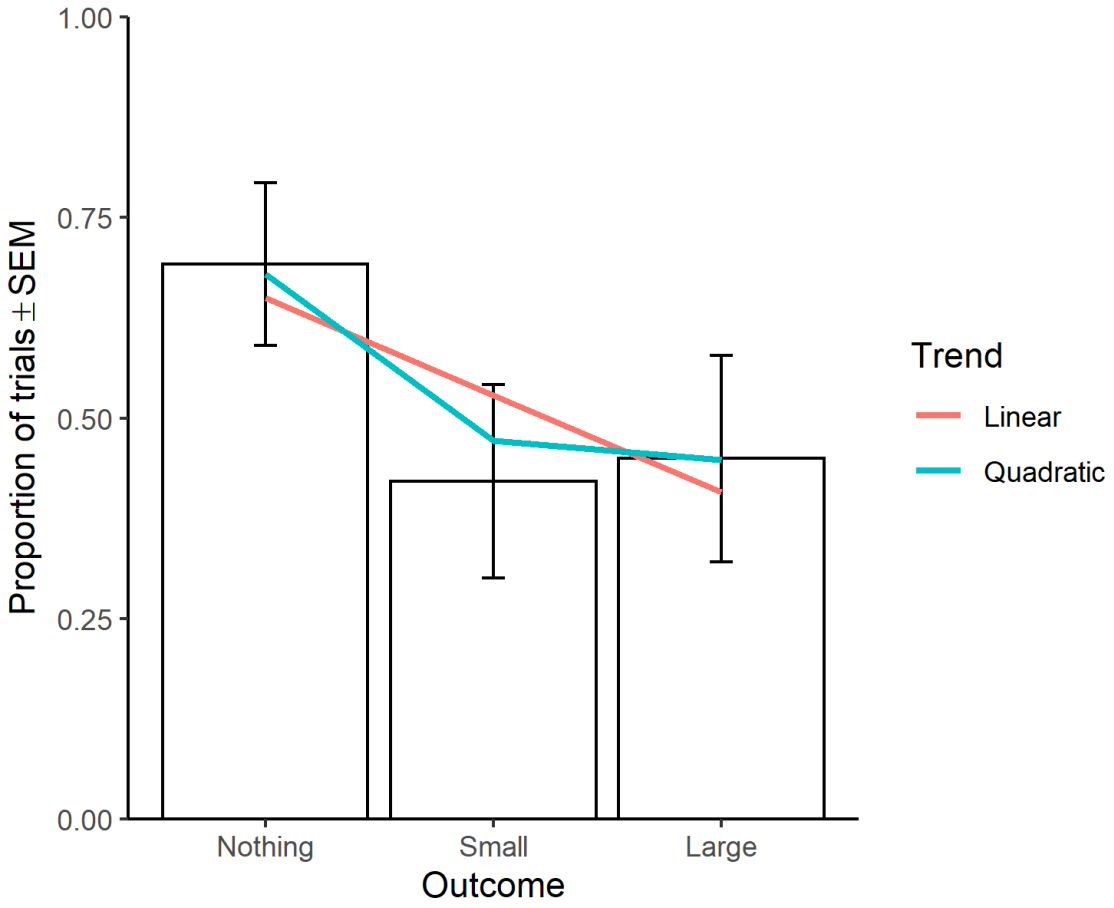


Figure 5.5.3: Proportion of trials in which subjects checked the unchosen tube as a function of outcome.

This represents a key divergence from counterfactual curiosity in children, who showed both a main effect of agency, checking the unchosen option more often in the agency condition, and an interaction with outcome, checking most frequently in agency trials where they received nothing (FitzGibbon et al., 2019). However, the two paradigms differ in two important ways. Firstly, the outcome for children was binary, either a match or not. Secondly, the use of glasses was mandatory and limited to one card, while in our study checking was optional and unrestricted.

Figure 5.5.4 shows the proportion of first looks directed at the unchosen tube as a function of agency and the binary outcome between nothing or something for only those trials where subjects checked at least one tube (n = 89). While the difference does not reach statistical significance (t-test, $t_6 = 1.93$, $p = .102$), we see that the relationship in the opposite direction, that agency plays more of a role in information seeking after positive than negative outcomes. While the small sample size in this study limits the possibility of finding small effects, this is an interesting pattern. Responding this way could be considered the most cognitive response, as it resolves the most uncertainty about the valence of the counterfactual outcome. While receiving an apple piece could reflect either regret or relief depending on the contents of the unchosen

tube, checking the counterfactual after a negative outcome can only lead to disappointment. A plausible further explanation for this is that participants are checking to confirm they received the highest value reward. The subjects rarely, if ever, checked the size of the apple piece before eating it, so it may be that checking for the absence of the large piece serves to confirm that one has answered correctly.

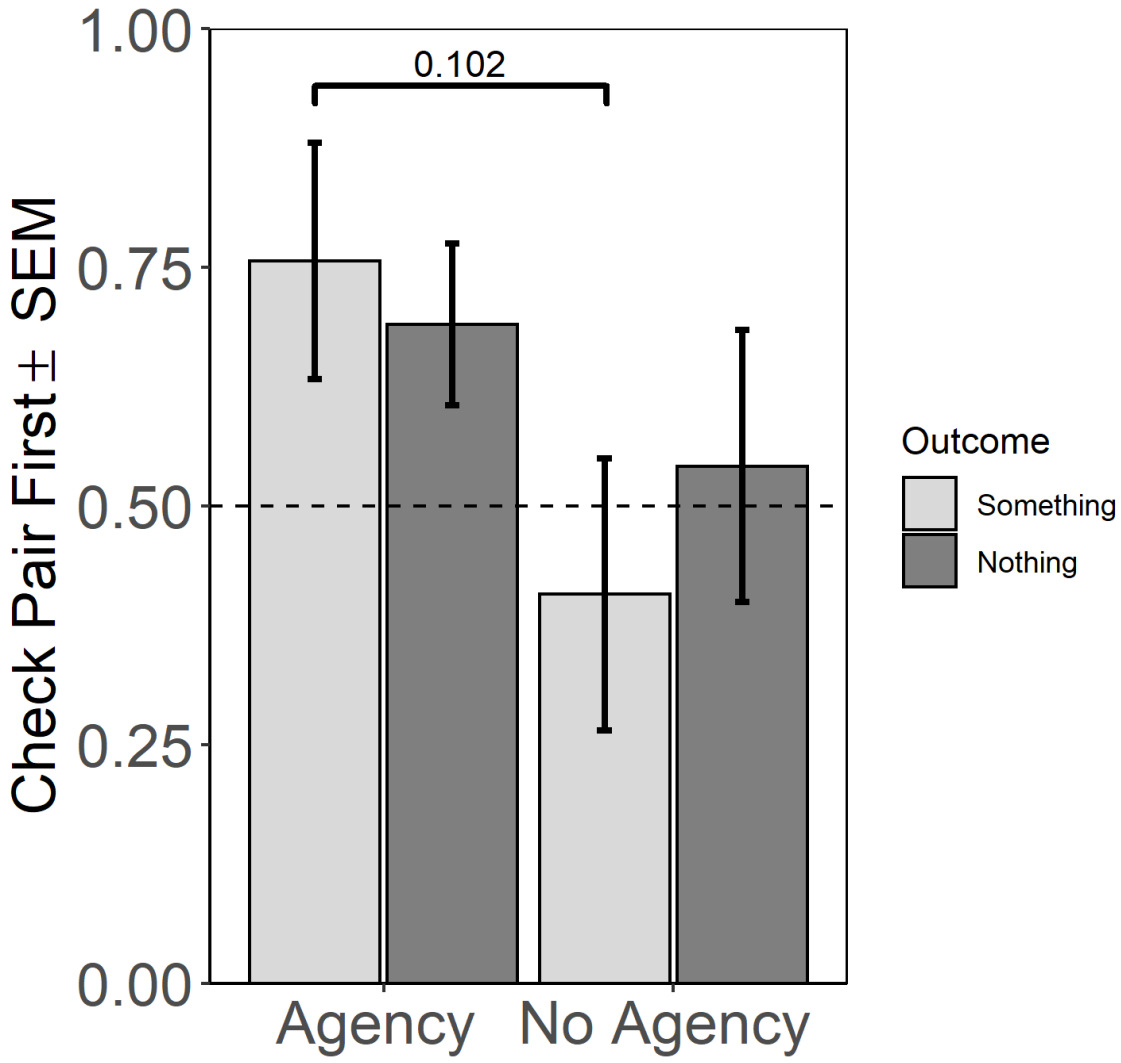


Figure 5.5.4: Proportion of trials in which subjects directed their first look at the unchosen tube, as a function of agency and outcome.

Table 5.5.2: Coefficients of a mixed effects model to predict the binary outcome of checking the unchosen tube in experiment 3.

	Term	β	conf.low	conf.high	p.value
(Intercept)	0.423	-1.139	2.039	0.606	0.544
outcome.linear	-0.998	-1.828	-0.225	-2.485	0.013
outcome.quadratic	0.764	0.048	1.525	2.066	0.039
No Agency	-0.458	-1.324	0.391	-1.066	0.287

5.6 General Discussion

Here we have demonstrated that chimpanzees are curious about what could have been, rather than just what is. Experiment 1 demonstrated that subjects bias their information search towards possible outcomes, rather than simply resolving the uncertainty as to what was in each tube. We further demonstrated that this was curiosity rather than information gathering, as subjects did not use this information in future trials. Experiment 2 tested a reductive explanation that subjects simply checked the tube that was closest to them, however the data do not support this conclusion. Finally, we manipulated whether subjects had agency over the tube which they received in Experiment 3 and found that the bias towards the unchosen tube was only present in the agency condition, but that there was no change in absolute rate of checking the unchosen tube between agency and non-agency conditions.

Unlike like the macaques tested by Wang and Hayden (2019), the chimpanzees aren't "paying" for the counterfactual information. In fact, Wang and Hayden's (2021) later operationalisation of curiosity includes that the information must not only have no instrumental value, but also the subject must show a willingness to pay that scales with the value of the information. Seeking counterfactual information in this experiment represented a minimal cost, simply moving one's head down to the level of the table. As discussed in the introduction and evidenced by subjects' willingness to check the (certainly empty) chosen tube in experiment 3, the cost to checking may be so low that it doesn't really represent an appreciable cost. To strengthen the claim that information-seeking is costly, future studies could place the tubes at ground level or above the subjects' head to increase the cost, while increasing the distance between the tubes may increase the selectivity of information seeking.

Wang and Hayden (2021) propose that the purpose of this information-driven counterfactual curiosity is to augment a cognitive map of the decision environment, suggesting a cognitive basis for the checking. In contrast, in attempting to answer why young children, who struggle with counterfactual reasoning, seemingly show a carryover from a learned association between counterfactual simulation and future utility, Fitzgibbon and Murayama (2022) suggest an effect

akin to an attentional trace. By their account, in the act of engaging with a choice between two alternatives, the focus directed towards the unchosen item leads young children to maintain an attention to it after their initial deliberation. This targeted information seeking is adaptive as it naturally drives an agent to investigate the alternate outcome without the need for mental simulation of a counterfactual, from which they could learn from their actions. Our data are not incompatible with this conclusion, particularly as we found no absolute difference in rates of checking the unchosen tube in the agency condition, only a bias towards it over the other tubes.

Although this does not fully explain the behaviour of macaques, who preferentially selected a gamble that provided counterfactual information, it is possible that the macaque data reflect uncertainty resolution, while the chimpanzee and child data may result from an attentional trace. Without a paradigm which tackles both elements we cannot say definitively. Future research could modify our paradigm to distinguish between these hypotheses by separating the deliberation and choice elements of the agency condition and measuring whether counterfactual curiosity persists.

If we do consider it in this manner, counterfactual curiosity is adaptive as it enables the optimisation of foraging strategies and the delineating of causal structures. So it may be that it is widespread within the animal kingdom, or equally, it could be a by-product of enhanced cognitive abilities unique to primates or shared with other highly intelligent species. Systematically extending this paradigm to more distally related taxa with varying cognitive abilities would hope to answer the question of whether this is a cognitive effect or a general attentional artefact. Finally, while our results are not a demonstration of counterfactual reasoning, they do provide the basis for its investigation and investigation into the counterfactual emotions.

6. General Discussion

6.1 Overview of the thesis

As humans we spend an awful lot of energy considering possibilities, whether that is planning for the future or pondering on the past, reasoning about what is possible is an essential aspect of our lives. For that reason, it is relevant to ask whether we are unique in our ability to consider multiple possible futures or pasts, and to reason which of these is possible. This thesis has been an exploration of how great apes reason about possibility, to answer the question of whether they do so in a manner comparable to our own reasoning, more similar to pre-verbal children, or wholly distinct. In this discussion I will recap the experimental paradigms that I used, then discuss the predictions from each model and whether the data support any of them. Following this I will integrate the data presented here with the wider literature covering three questions: why performance breaks down in the 4-cup task relative to the 2-cup task; whether inference ability exists on a spectrum; and whether the type of uncertainty matters for how great apes reason about possibility. Finally, I will give my perspective on whether language is a pre-requisite for logical reasoning.

Introduction

In the introduction I discussed firstly how adults consider uncertainty and how we can use different forms of reasoning to confirm, rule out or adjust our expectations of what is possible or probable. Secondly, I discussed two theories from the developmental literature, the minimal model (Leahy & Carey, 2020) and the temporal junctures model (Redshaw & Suddendorf, 2020), both of which suggest that reasoning about possibilities is unique to humans. As they make similar predictions, I will simply refer to both as the *minimal models* and specify when the temporal junctures model diverges. Subsequently, I discussed comparative literature which conflicts with these models, and alternative theories which have come from a comparative perspective. From this overview we can see that none of the theories are sufficient to explain all of the findings in the primate literature

Chapter 2: What is possible?

In Chapter 2, I introduce a novel addition to a basic 2-cup paradigm, post decision wagering. Post-decision wagering asks subjects to rate their confidence in their decision retrospectively, by offering the subject a revised choice between their original selection and a certain piece half the size of the original. The dependent variable is the individual rates of taking the half grape

between conditions. The paradigm was designed to directly test the minimal models, which suggest that non-human animals are only able to prepare for a single uncertain outcome. In the first and second experiment we tested for differences between a certain outcome, where the subjects saw where the whole item was placed, and an uncertain outcome, where the baiting took place behind an occluder. In the final experiment, both conditions were baited behind an occluder, but the conditions differed by the number of cups behind the screen during baiting, one in the certain condition but two in the uncertain condition.

Under the minimal models, we should see no difference between the rates of taking the half-piece in the certain and uncertain conditions, because, if the subject has only simulated one possible scenario, they should treat the certain and the simulated outcomes the same, which the data do not support. However, this only applies to the strictest reading of the minimal model, alternatively, if subjects are either recognising cues to uncertainty (such as the occluder) or making the choice based on differences in the whole representation then this could explain opting out of the occluded trials. To counter this, experiment 3 matched the strength of the overall representations by baiting both conditions behind an occluder, as we still report a difference between the one-cup and two-cup conditions, we can reject the minimal models.

Under the ratio of ratios (RoR) account (Hanus & Call, 2014), the probability estimation system used by apes is unable to discriminate $p = 1$ from $p = 0.5$, so subjects should therefore treat both equally relative to the half-grape, which they do not. Once again, in experiments 1 and 2, structural differences between the conditions may have meant that apes weren't estimating probability in the visible condition, because they had seen the food hidden. In the third experiment, when we controlled for this by matching the presentation between conditions, we continued to report a difference between conditions, so we can reject the RoR account.

The location-based argument (Engelmann, Haux, et al., 2023) would not make strict predictions about visible trials of the first two experiments; in occluded trials however, if apes were truly ambivalent to their likelihood of finding the target item when more than one cup remains, then it would be logical to continue searching the broad location which the whole grape is certainly still under, half of which is still available to them. However, we do not know whether the large location becomes half as attractive when only half of it is in reach. Like the RoR account, in Experiment 3, the location-based explanation predicts the subjects treating these two as equivalent, which they do not.

The probabilistic model, which generally attributes rational behaviour to non-human animals, would predict behaviour in line with expected value, meaning that subjects wouldn't take the half piece on any certain trials, and responses on uncertain trials would be based on individual

risk preferences, as the expected value of the two choices are equal. Of all of these theories this is the only one which we find support for, which suggests that apes do have a rational response to epistemic uncertainty.

Chapter 3: 2-cup Inference

In Chapter 3, we continued to apply this paradigm of post-decision wagering but instead aimed to investigate whether apes could use the disjunctive syllogism to reason about the contents of the unchosen cup if we gave them information about the contents of the unchosen cup. Leahy and Carey (2020) propose the disjunctive syllogism as being evidence of modal reasoning, finding that apes are able to reason via the disjunctive syllogism would demonstrate that language is not a pre-requisite for its emergence. Gautam, Suddendorf and Redshaw (2021a) have argued that reasoning via the exclusive disjunction, A or B, not both, is more complex than reasoning via the inclusive disjunction, A or B, maybe both, because it requires an additional negation. Baboons and children between the ages of 2½ and 5 are between these stages (Ferrigno et al., 2021; Gautam et al., 2021b). If apes were also at this stage, we would expect to see them perform well in trials where the unchosen cup was empty, and not take the half piece; but in trials where the target piece was removed subjects would continue to choose at rates equivalent to the no information condition from Chapter 2.

The location-based explanation would propose ceiling performance in both conditions, because in the reveal empty trials subjects would simply shrink the location to include only the chosen cup, while in reveal baited trials they would remove the location entirely and select the half grape. Likewise, the probabilistic model would also predict rational performance. Under the avoid the empty cup hypothesis, we would expect to see no difference in rates based on the contents of the unchosen cup, as the reasoner does not make expectations as to contents of the chosen cup, they simply avoid the now-empty unchosen cup, which was not available to choose in either condition.

In Experiment 1 the data are mixed, overall, the group takes the half-grape more frequently in reveal baited than reveal empty trials, however not significantly so, meaning that we cannot reject the minimal models or the avoid the empty cup hypothesis. Yet when we examine the sessions separately, we see that they initially responded differentially based on the contents of the revealed cup before reverting to chance in the second session, a behaviour that is difficult to explain. At an individual level, two individuals switched adaptively based on the contents of the unchosen cup, one of whom scored 100% in both trials. Therefore we find support for the conclusion reached by Schmidt and Fischer (2009) for baboons, that some individuals are

capable of inferential reasoning, while others are not. This however, goes against previous literature in great apes (e.g. Call, 2006; Engelmann et al., 2022) including when controlling for empty cup avoidance (Call, 2022). Contrastingly, in Experiment 2 the Twycross apes, who were inexperienced with cognitive testing, did adapt their switch rates based on the contents of the unchosen cup, which suggests that for this group we can reject the minimal models. Additionally, the group adapted their choice rates in both conditions, which suggests that they are equally able to resolve the exclusive and inclusive disjunction, this is true whether you test them against chance, or against their individual rates when the unchosen cup wasn't revealed (Chapter 2)

In Experiment 3, with a modified paradigm which aimed to eliminate the possibility of passing the task by stimulus enhancement, we find that the Edinburgh group did switch differentially, and all individuals adaptively altered the half-choice rates in the predicted direction. As such, it would appear that their reversion to chance in the second session of Experiment 1 was an anomalous result. Additionally, in Experiment 3, the Edinburgh chimpanzees switched at a rate comparable to the certain trials of Chapter 2 Experiment 1, which suggests that they are treating their inferences as equivalent to visual information, suggestive of *deductive* reasoning. The Twycross group however did not, the reasons for this are unknown, but it could be down to an experienced-based increase in reliance on inferences in the Edinburgh group. Equally, it could be a task factor such as the miming of the action drawing attention to the absence of a food piece.

Therefore, we can reject the minimal models of possibility, and avoidance of the empty cup. While performance in reveal empty trials was close to ceiling, it was not for the reveal baited trials, so we should also reject a strict reading of the location-based hypothesis. For both experiments 2 and 3, performance in both conditions was significantly above chance, so we can also reject the conclusion that apes are viewing the OR as an inclusive disjunction. Nonetheless, there is a spectrum of individual differences within the Twycross group, so it is entirely possible that some individuals do fail to represent its *exclusive* meaning.

As in Chapter 2, this rejection of all simpler hypotheses suggests that the data support the probabilistic reasoning hypothesis. It would appear that there is a heavy discounting for decisions reached via inference, which would have the effect of moving both conditions towards an individual's base rate. This could also be used as an explanation for the difference between the experienced Edinburgh group and the naïve Twycross group, if the Edinburgh group have a lower discounting for inference, on account of their increased experience, then we would expect to see a greater difference between conditions.

Chapter 4: 4-cup inference

In Chapter 4 we present the original (Mody & Carey, 2016) and the modified (Ferrigno et al., 2021) 4-cup paradigm to the Twycross group. In the two variants, one cup is revealed, either because of the subjects first guess (modified) or as part of the experimenter's demonstration (original), allowing the subject to infer the contents of its pair. The dependent measure is the rate of switching to the uncertain pair, with chance set at 0.67, as it makes up two of the remaining three cups.

Under the empty-cup avoidance hypothesis (Paukner et al., 2006), apes do not represent the likelihood of each cup containing a food item as dependent on the contents of its pair, so the content of the revealed cup is irrelevant, so we should see no difference between any of the conditions and all choices should be random.

Under the minimal model, ratio of ratios and location-based approaches we would predict 50% performance in reveal empty trials, as subjects do not discriminate between the certainly full and the potentially full cups, so simply pick randomly between the pair and the target cup; and ceiling performance in the reveal-baited trials, because the first piece has been removed. While not addressed directly, my understanding is that the RoR would also predict high performance, as the subjects are able to infer the cup is empty it is no longer an uncertain choice, so there is no RoR to calculate [28]. Conservatively, we would expect performance equal to the reveal baited trials of Chapter 3.

Under an *inclusive only* understanding of disjunction (Gautam et al., 2021a), we should find the same results as Ferrigno et al. (2021) and Gautam et al. (2021b), that subjects are above chance in the reveal empty condition, but precisely at chance in the reveal baited condition. Finally, the probabilistic approach would predict high performance in both trial types with a consistent rate of error, ascribed to the chance of a false positive from the contents of the revealed cup, for the visual system this would be expected to be very low. Notably, none of these theories from the literature would predict differences in performance between the two variants of the paradigm.

In both variants of the task, we see a difference based on condition. On average subjects switched on 52.5% and 64.5% of reveal empty and reveal baited trials, respectively. Therefore, we can reject the empty cup avoidance hypothesis and attribute at least some level of inference,

[28] If we did attempt to it would be infinite as the RoR is calculated as the relative likelihood of getting a desired outcome from the more favourable choice (Eckert, Call, et al., 2018), this would mean dividing by zero, therefor an infinite RoR.

as apes are responding adaptively to the contents of the revealed cup. But must also reject the strict probabilistic account, as apes are failing to reason correctly on almost 50% of trials, although changes to the rates of switching in each condition were in an adaptive direction. In both the one-choice and the two-choice variants this divergence is driven by the performance in reveal empty trials, which was above chance, while switch rates in reveal baited trials were not. This would suggest, as Gautam, Suddendorf and Redshaw (2021) did with regard to the baboons tested by Ferrigno et al. (2021), that apes are able to resolve the inclusive disjunction, but fail to comprehend the exclusive disjunction.

Notably, while performance in reveal empty trials is above numerical chance, it is not significantly above the 50% that would be expected by the minimal models, thus the data do not reject those theories. However, the performance in reveal baited trials does not conform to the minimal models, if the minimal agent simulated the location of both food pieces at the start of the trial, then when one was revealed, they should have switched to the other for their next guess, leading to 100% switch rates. Instead, we see subjects choosing randomly for their second guess. Which also means that we should reject the location-based and RoR accounts. Resultingly, the suggestion that primates and young children treat the disjunction as inclusive (Gautam et al., 2021a), is the best explanation of our data.

In the control task, we retested those individuals who had scored above chance in at least one condition of the standard task alongside the Edinburgh chimpanzees, but manipulated whether the pairs were baited as standard, or both food pieces were placed into one pair. The intention of this manipulation was to investigate whether subjects were able to flexibly apply the logical operation they had used in the first experiment or were simply deploying an associative strategy of switching pairs when they found a piece and staying if they did not. The results we found were inconclusive. Firstly, we found that there was no significant difference in error rates between the two-groups, which suggests that the successful Twycross apes were not using an associative strategy. However, performance was far from perfect in either condition, on approximately 50% of control trials (where both pieces were placed into one pair of cups), apes chose an unbaited cup, despite the unbaited pair never being placed behind the occluder. While the exact reason for this irregular behaviour is unresolved, the experiment concluded that apes were not able to flexibly apply the logical operation required to pass the 4-cup task, suggesting that they were not using logic to solve it.

What was possible?

In Chapter 5 we change our focus and instead explore whether chimpanzees are curious about counterfactuals. Outcomes which were at one point possible, but now are not. Although the other models do not make predictions as to whether primates can consider counterfactuals, under the minimal models, non-human animals are concerned simply with what is, not what could have been. However, the data show that chimpanzees are driven to investigate the outcome of the option which they could have picked but didn't, over one that was never available to them. While this is only a demonstration of curiosity and not full counterfactual reasoning, it is not supported by the minimal models.

Nevertheless, as proposed by Fitzgibbon and Murayama (2022), this could be a simple attentional artefact, as there is no reason for young children, who cannot yet reason counterfactually, to exhibit a carryover between counterfactual reasoning and intrinsic value. They argue that, despite not eventually settling on it, the attention which the subject paid to the unchosen option may lead their attention to be drawn to it subsequently. Future research could investigate whether this attentional trace is modulated by the level of information available from seeking the counterfactual. To utilise a paradigm from inference research (Jelbert et al., 2015) if the unchosen tube has a 90º bend in it, so does not provide any information when checked, will the attentional trace still drive chimpanzees to check it after the trial? If not then it may suggest that chimpanzees are, in-fact, seeking information from the unchosen outcome. While the cognitive underpinnings of the data are still to be resolved, we continue with a discussion of whether counterfactual simulation could be possible and lay the groundwork for an investigation into counterfactual emotions such as regret and relief in non-human animals.

Consolidation.

Firstly, to investigate whether apes were able to consider multiple possibilities, we added a post-decision wager to a basic 2-cup task and found that apes were able to adaptively adjust their rates of taking a fractional piece as we varied the level of uncertainty in the task (Chapter 2). When we then gave subjects indirect information about the contents of their selected cup, they then adapted their rates of taking the fractional piece in line with reasoning via the disjunctive syllogism (Chapter 3). The data presented in the first two experimental chapters allowed us to reject several reductive explanations for why primates had previously performed poorly in the 3- and 4-cup tasks (Engelmann, Haux, et al., 2023; Gautam et al., 2021a; Hanus & Call, 2014; Leahy & Carey, 2020; Paukner et al., 2006; Redshaw & Suddendorf, 2020) . Instead we find support for a probabilistic account (Rescorla, 2009), which broadly ascribes rational behaviour to non-human animals, but falls short of full deductive reasoning.

In contrast, in the following chapter, when we tested these same individuals again under both the original (Mody & Carey, 2016) and the modified (Ferrigno et al., 2021) 4-cup paradigms, we found that they were unable to effectively deploy the same logical operation. We find that while apes did adaptively switch between pairs based on the revealed cup, allowing us to reject the minimal models, overall performance was poor (~50%). Testing against chance showed us the condition effect was driven by the reveal empty trial type, supporting an inclusive only understanding of the disjunctive (Gautam et al., 2021a) rather than the complete understanding demonstrated in the preceding two chapters. Moreover, when we added in additional trials to test for flexible application of the logical operation, adaptive switching broke down completely, and subjects were prone to selecting cups that hadn't been behind the occluder, so could never have been baited. Finally, in chapter 5 we showed that chimpanzees were motivated to investigate unrealised outcomes, suggesting once again that they do consider multiple possibilities.

Overall, the data suggest that the minimal models are not suitable descriptions of how great apes reason about possibility but that 4-cup task is not an effective means for investigating disjunctive reasoning in great apes. I propose that there are additional factors which are taxing subjects' cognitive capacities, and, because performance breaks down in different ways, developmental and comparative literature should be considered separately. As such, future explanations should instead focus on what factors are causing the breakdown in performance between single item inference tasks and their multi-item equivalents.

6.2 How does the breakdown in the 4-cup task relate to theories of human reasoning?

The data presented here conform with the literature that great apes can accurately reason in a 2-cup-1-item task, but that reasoning breaks down in the more complex tasks, despite both relying on the same underlying logic. Under a *standard logic* account (Rips, 2001) we apply logic sparingly to evaluate focussed parts of our model. This means that in the 4-cup task one can simply address the pair which was acted on by the experimenter, to calculate the contents of the unmanipulated cup and choose between it and the pair. Under this model there would be no breakdown of reasoning in the 4-cup task, because it is iterating the logical operation of the 2-cup task over multiple pairs, so this cannot explain our data.

Alternatively, the *mental models* approach to reasoning suggests that when we are faced with a reasoning task, we populate all of the possible models, we then go through a sequential process

of winnowing until we are left with only what is possible (Johnson-Laird, 2010). Notably, the mental models approach also relies on logic to rule out possible models, but the difference is that it requires reasoners to consider all models of the scenario in parallel rather than applying logic selectively to focussed elements. Under this explanation, at the start of the trial the reasoner has 4 possible models for where the grapes could be in cups A to D: AC, AD, BC, and BD. If the experimenter shows them that B is empty, they rule out the second 2 models, so the grapes must be in either AC or AD. It then takes a *rational* agent, to behave in a way that is most adaptive to their wellbeing (Johnson-Laird, 2021). Evans (2021) makes a distinction between being *functionally* rational, taking the option that is in an agent's best interest, and *epistemically* rational, holding beliefs that are the most accurate. In our context, it takes both an epistemically rational agent to maintain an accurate model of the likelihood of each cup containing a food piece and a functionally rational agent to select the cup with the highest likelihood.

From a mental models perspective, apes fail the 4-cup task either because they select the wrong cup after forming an accurate model, or fail to populate an accurate model of the (likely) contents of each cup, but choose the cup they *think* has the highest likelihood. The conclusion that apes are not functionally rational is summarily countered by data from chapters 2 and 3, in which they choose in line with expected value. Of particular significance is the third experiment of chapter 2, when the level of uncertainty was dictated by the number of cups behind the barrier during baiting. As measured by their relative rates of taking a constant half-grape, chimpanzees differentiated between a cup that certainly contained a grape and one that only possibly contained one, which is precisely the distinction required in the 3-cup task or reveal-empty trials of the 4-cup task.

Consequently, under a mental models account apes must fail the 4-cup task due to being epistemically irrational, possibly due to failure to maintain separation between the models of two functionally equivalent 2-cup tasks. Plausibly, if we were to offer a sequential version of the 3- or 4-cup task, in which the inference operation on the first pair was resolved before the second pair was offered, we may see an improvement in performance. We do know that 14-month-old children, who are well below the age at which children pass the 3- and 4-cup tasks, are able to accurately track small sets and keep these representations separate (Rosenberg & Feigenson, 2013). Meaning that, at least for children, the mental models account cannot account for this failure. Nevertheless, divergence between children and apes on reveal baited trials of the 4-cup task (Engelmann, Haux, et al., 2023; Gautam et al., 2021b) suggest that there are different task elements limiting performance.

A second school of thought is that humans are capable of being rational, but we simply don't devote all of our mental effort to the problem at hand. Instead, we have two competing 'systems', *System I* and *System II,* which we use for thinking hence this is termed the *dual-systems* model(Stanovich, 1999)[29]. For day-to-day tasks we use System I, System I is fast, does not require excessive cognitive effort and relies on associations, but is prone to making mistakes; System II, however, is the opposite, it is slow and effortful, but rational and doesn't make mistakes. For this reason, popularised versions of a dual-systems approach have argued that System II 'takes over' from System I only when it is required, thus explaining why we often make mistakes in our reasoning. This also fits somewhat into the modular brain hypothesis (Tooby & Cosmides, 1992), System I is specialised and modular, so learns associations specific to a task, while System II is more like a general intelligence applied to lots of different tasks, Stanovich (2016) refers to this as a reasoning quotient.

Evans and Stanovich (2013) discuss the improper application of dual-systems models, instead opting to specify Type I processes and Type II processes, to refer to fast, subconscious thinking and slow effortful thinking, respectively, and reflect on the types existing on a sliding scale. In the cognitive tasks which we present to the apes, subjects receive 8-12 trials per session in quick succession and for each they receive only a small food-reward. It is entirely possible that the cost-benefit ratio of engaging with Type II thinking is not high enough for the task at hand. Potentially, in future studies which utilised a higher value reward and a longer pause between the reveal and choice, a more protracted choice process would result in more rational responses. Mercier and Sperber (2020) argue that even reasoning is a Type I process, citing the evidence that when our reasoning is flawed, we often accept our own justifications solely on the fact that they came to us quickly – the availability heuristic. So, it's possible that this time delay would make no difference. In their own theory of human reasoning, Mercier and Sperber (2011) place language as essential in the development of proper reasoning and believe that it is only the anticipation of being challenged which leads us to effective reasoning.

Gigerenzer (2011) takes aim specifically at Tversky and Kahneman's heuristics and biases programme. He argues that heuristics are not a negative which should be lumped in with biases but instead an asset which can aid in making fast and rational decisions with limited information. He argues that there is no infallible System II whose purpose is to correct System I. Instead, we are bestowed with an adaptive toolbox (Gigerenzer & Todd, 1999), a set of adaptive

---

[29] Stanovich (1999) used 'system' as a neutral term to show that he was agnostic to the different names in use at the time, rather than intending to declaring these as two separate systems (J. S. B. Evans, 2006), however modern theories popularised the 'systems' approach as described here (Kahneman, 2011).

modules, each of which serves a purpose and can be used either consciously or subconsciously. For the same reasons as mentioned above, these cognitive tasks may be an ideal place for the development of heuristics, such as the *win-switch lose-stay* strategy tested in Chapter 4. The generalisability of these toolbox elements is important however, much like the selective application of logic, applying a heuristic decision rule to a pair of cups in a 4-cup task should be no different to applying it to a pair of cups in a 2-cup task. Therefore, none of the human theories of reasoning breakdown can conclusively explain the performance of apes in the 4-cup task, and paradigms should be designed to explicitly test them.

6.3 Does inference ability exist on a spectrum, are individual differences consistent?

From this set of experiments have emerged a set of "high performers". From the Edinburgh group neither Frek nor Velu selected the half grape on any reveal empty trials in Chapter 2. In Chapter 3, Velu's performance was exceptional in both experiments, while Frek also switched adaptively in both. Although anecdotal, in the pre-test of the counterfactual curiosity experiments, if only nothing or the small piece were available in the uncovered tubes, Velu often pointed to the 3rd (still covered) tube, having inferred that it contained the larger reward. However, in the 4-cup task, neither Velu nor Frek were above chance in the standard or the control conditions.

From the Twycross group, it is only Likemba who performed well in both the 2-cup and the 4-cup task, and at the group level there was no correlation between the two. Kibali scored comparably to Velu in the 2-cup task but also was at chance in the 4-cup. Interestingly, he scored very well in the first session of the 4-cup task, albeit with a strategy of always initially selecting cup 3 then 2 or 4 depending on the outcome, then reverted to chance in the following sessions[30]. Kayan, who was above chance in both reveal empty and reveal baited trials of the 2-choice 4-cup task scored very poorly in the 2-cup task, as did Lope. It is worth remembering the Twycross apes were almost completely inexperienced to face-to-face cognitive testing, the studies presented here represent the first exposure to object search tasks for the gorillas and orangutans and only the second for the bonobos and chimpanzees. While they do receive cognitive enrichment that involves hidden food, the rest of the process, including the

[30] Kibali's second session took place on a day when there had been a large fight within the group, although he had not been involved, this could have taxed his attention. In turn, poor performance may have led to acceptance that the task was too challenging and that answering at random was the best strategy.

requirement that they inhibit selecting a visible food item, was novel. Both Kayan and Lope took the half grape on approximately 2/3 of 2-cup reveal empty trials yet passed the 4-cup reveal empty trials at above even the 50% chance level. It is possible that if were to go back and re-test the 2-cup task, their performance would be better. Or equally, inhibiting selecting a visible piece may continue to be a challenge for them and their result may be no different.

Interestingly, several individuals did well in one task but did not participate in the other. Basuki a 5-year-old male orangutan and his mother, Maliku, performed well in the 2-cup task but did not participate in the 4-cup due to the adult male monopolising access. Lola, a 3-year-old bonobo, did very well in the 4-cup task, but was not yet independent of her mother, Likemba, during the 2-cup testing phase. These individuals may have tested well in both, but we also have no specific reason to believe that their scores across experiments would correlate while most others did not. Thus, we would conclude that there is no support for a general inference ability, in contrast to the suggestion by Herrmann and Call (2012).

There are two possible explanations of this clear divergence between two tasks ostensibly testing the same ability, either the individuals who have performed well in one task have happened upon effective decision rules specific to that task and are not using inference at all, or secondly, as I have argued throughout, that performance in the 4-cup task is limited by a different non-inferential factor.

6.4 Does the type of uncertainty matter?

In the original forked ramp task (Beck et al., 2006), 3- and 4-year-old children found it more difficult to prepare for 2 mutually exclusive possibilities than they did to answer questions about a hypothetical counterfactual after the fact. Similarly, children between the ages of 4 and 8 found it easier to prepare for an undetermined event (physical uncertainty) than a comparable epistemic event, where they were simply ‘hedging their bets’. This would suggest that epistemic uncertainty places the highest demands on the thinker, thus, counterfactual thinking and further physical uncertainty tasks may be possible in non-human primates. There have been two proposed reasons for why this might be, either that being asked to reflect on their own knowledge is cognitively taxing so detracts from cognitive capacity (Robinson et al., 2006), or because it requires the thinker to treat an expired temporal junction as though it were live (Redshaw & Suddendorf, 2020), thus the same cognitive requirements as counterfactual reasoning but without the need to negate the real world. On a separate note, this final point could be why counterfactual reasoning emerges later than reasoning about epistemic uncertainty.

I would argue that a non-linguistic reasoner does not necessarily need to think about epistemic uncertainty in either of these ways. If instead we simply consider Friston's (2010) view of uncertainty: that the agent has a model of the world, and some aspects of that model are more fuzzy than others. The agent does not need to reflect on the fact that their model is incomplete, as many other aspects of their model are incomplete, the contents of the cups are just one aspect of it. This would mean that an agent could respond to epistemic uncertainty without reflecting on their knowledge state, or, as proposed by Gautam, Redshaw and Suddendorf (Gautam et al., 2021a), conducting mental time travel to reason about the events which could have led to the food being in each of the two locations. Crucially, any element of the model which you don't have current visual access to would have an associated level of uncertainty, ranging from low in the case of an item you had observed being hidden, to high for an item whose location is genuinely unknown. It is then possible to apply the RoR account to the likelihood of success for finding each item. I would hypothesise that if the contents of the certain cup were visible in the original 3-cup task (Hanus & Call, 2014), then we would not see the 50% choice rates.

Physical uncertainty on the other hand cannot be solved using a static model, as subjects must simulate into the future to reason about the outcome of events that have not happened yet. However, a recent study has shown that chimpanzees are also able to prepare for multiple possibilities under physical uncertainty (Engelmann, Völter, et al., 2023). The authors replicated the forked tube paradigm (Redshaw & Suddendorf, 2016) but in a competitive rather than a cooperative setting, as this has previously been a more effective approach for chimpanzees to express their intelligence (Hare et al., 2001). In the revised paradigm (Engelmann, Völter, et al., 2023), a second experimenter dropped a stone into a forked tube to attempt to dislodge two trays baited by the first experimenter. If the chimpanzee stabilised the trays, then when the original experimenter re-entered the room, they would give the subject the food which remained on the trays. Subjects stabilised both trays more frequently in the forked tube condition than a control, in which a single straight tube was positioned over one of the trays, allowing the authors to argue that they have prepared for both possible outcomes.

The trays paradigm is a valuable development as it is open to a variety of follow-up studies to further develop our understanding of how great apes consider different types of uncertainty, for example the socially determined version with two straight tubes (Suddendorf et al., 2017), preference attribution as in the urn task (Eckert, Rakoczy, et al., 2018) or using physical barriers

(Crimston et al., 2023) to block the tube from posing a threat to the trays[31]. Moreover, if we were to combine the paradigms deployed in this thesis, such as testing whether great apes could plan for multiple eventualities by moving their body into a position where they could see both outcomes before they occurred, this would provide a competitive context and demonstrate a form of anticipatory metacognition.

From the other perspective, it could be said that apes are considering epistemic certainty as children are, so their successful reasoning about it means they could theoretically be capable of counterfactual reasoning and counterfactual emotions. As discussed earlier in the thesis, counterfactual reasoning requires that subjects make causal inferences within a simulated world, while disregarding the actual state of the world. Firstly, there is ample evidence that apes and other species are adept at making causal inferences (Völter & Call, 2017). Secondly is the requirement to make simulations. This is a cognitively demanding ability alongside being exceedingly difficult to demonstrate. However, a rich literature on hippocampal simulation, both forwards and backwards, exists in rodents (Comrie et al., 2022; Mahr, 2020). While this evidence for simulation is convincing, behavioural evidence answering the question of whether these agents are aware of these simulations is currently incomplete (c.f. Redish, 2016). On a behavioural level, the rich interpretation of 'trap' tasks, in which subjects choose to push a food item either left or right in a tube to avoid it falling into a trap, requires that subjects simulate their actions ahead of time. Both corvids (Seed et al., 2006) and apes (Mulcahy & Call, 2006; Seed et al., 2009b; Völter & Call, 2014b) have passed these tasks, which would imply that forward causal simulation is within the capacity of these species. However, whether these truly resemble the human capacity for autonoesis, to place oneself within a mental simulation, is a more complex question.

Furthermore, the developmental literature appears to show that forward thinking may be simpler than full counterfactual reasoning, for example, children consistently find it easier to answer questions about future alternative scenarios than comparable counterfactuals (Beck et al., 2006; Perner et al., 2004; Robinson & Beck, 2000). The additional cognitive complexity placed on the counterfactual thinker by the negation of the factual world, a crucial delineating factor between real-world- and general counterfactuals, has been proposed as an explanation for the former's

[31] One possible non-cognitive explanation for the behaviour is that during the observation phase chimpanzees learned a rule that "trays under tube openings sometimes get knocked off", so stabilised them. While I don't believe this is the true cause of the behaviour, great apes have shown previously to fail to comprehend the causal significance of tubes (Cacchione & Call, 2010), and these follow-ups could control for this explanation.

delayed developmental emergence (Beck & Riggs, 2014). Hence, this could feasibly make real-world counterfactual reasoning uniquely human. Nevertheless, Chapter 5 has demonstrated that chimpanzees actively search for information about unrealised outcomes of their own choices, and while this does not necessitate counterfactual reasoning, it does provide a basis from which to investigate counterfactual reasoning and counterfactual emotions in non-human animals.

6.5 Conclusion: Is language necessary for logical reasoning.

From a human perspective, the interactionist theory of human reasoning (Mercier and Sperber 2010) suggests that reasoning originally evolved for a communicative purpose, as an interactive strategy to critique the merits and flaws of different plans and strategies. Subsequently, individual reasoning emerged as an introspective simulation of inter-personal reasoning, and when faced with conflicting theories we will gravitate towards the one which is easiest to justify (Mercier & Sperber, 2011). The argumentative theory places language as an integral aspect in the emergence of reasoning, but crucially, individual reasoning acts on intuitions formed from environmental regularities, which are available without the power of language. These are the building blocks on which inference takes place, but unlike language they are available to non-human animals. On an individual level, either the recognition of these environmental regularities, or the ability to abstract them out of the specific scenario in which they occurred could be the basis for the individual variation which we have reported. However, the lack of generalisability between the 2- and 4- cup disjunctive syllogism task suggests that cross-task transfer may be limited.

The language centric definition is reminiscent of the distinction between inductive and deductive inference, both of which share the end-goal of using held knowledge to derive new knowledge, where they differ from one another is the method by which they reach the new knowledge. Deductive reasoning uses strict rules or axioms, statements which are known to be universally true, to make new prepositions from initial premises; induction, by contrast, does not require formal rules and instead works via drawing logical conclusions based upon previous observations (Henderson, 2020). The characteristic difference being that deductive inference relies on absolute truths, while inductive inference relies upon probabilities.

While non-human animals may be capable of tracking these environmental regularities, having intuitions, and making logical choices based on them; these conclusions are likely still probabilistic, as deductive reasoning is an entirely language-based concept. While a notable peril of experimental primate cognition research is that subjects are exposed to various paradigms over time, thereby endangering carryover between tasks, this may only place them on

a par with the exposure that a developing infant receives, thus allowing them the experience to make the abstract rules ubiquitous in human development. Nevertheless, if we return to the definition of logic from Watson et al. (2001), *"…each failure to find the object in a selected place amounts to an increase in the implied likelihood of the object being at a place not yet searched*", then the apes we have tested do conform to the tenets of logical search. Therefore, it is reasonable to conclude from this thesis that search behaviour of non-human great ape is in fact guided by logic, just not deductive logic in the human sense.

Grigoroglou and Ganea (2022) point to the polysemy of many of the modal verbs and argue that children learn the simpler non-epistemic uses of the word first, which they then use to scaffold the modal concept later. The authors note that children do not start to use the semantic meaning of the word in an adult sense until the age of 7, it is plausible that the young children tested by Mody and Carey (2016) and by Gautam, Redshaw and Suddendorf (2021), are also responding simply with intuitions. However, the question remains as to why reasoning via the disjunctive syllogism under the 4-cup paradigm is near-ubiquitous in children by the age of 5, but only present in a handful of apes tested. As I have argued throughout the thesis and supported by evidence from chapter 3, the possibility which always exists in comparative research is that there are task constraints which are limiting performance, and redesigned paradigms or testing behaviour indirectly may provide more answers. To this end, conclusively testing deductive inference using a visual search paradigm may not be possible and better answers could come from novel physiological measurements that can characterise violations of expectation.

7. References

Amici, F., Cacchione, T., & Bueno-Guerra, N. (2017). Understanding of object properties by sloth bears, *Melursus ursinus ursinus*. *Animal Behaviour*, *134*, 217–222. https://doi.org/10.1016/j.anbehav.2017.10.028

Banerjee, K., Chabris, C. F., Johnson, V. E., Lee, J. J., Tsao, F., & Hauser, M. D. (2009). General Intelligence in Another Primate: Individual Differences across Cognitive Task Performance in a New World Monkey (Saguinus oedipus). *PLoS ONE*, *4*(6), e5883. https://doi.org/10.1371/journal.pone.0005883

Barth, J., & Call, J. (2006). Tracking the displacement of objects: A series of tasks with great apes (Pan troglodytes, Pan paniscus, Gorilla gorilla, and Pongo pygmaeus) and young children (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(3), 239–252. https://doi.org/10.1037/0097-7403.32.3.239

Bault, N., Wydoodt, P., & Coricelli, G. (2016). Different attentional patterns for regret and disappointment: An eye-tracking study. *Journal of Behavioral Decision Making*, *29*(2–3), 194–205. https://doi.org/10.1002/bdm.1938

Beck, S. R. (2016). Why What Is Counterfactual Really Matters: A Response to Weisberg and Gopnik (). *Cognitive Science*, *40*(1), 253–256. https://doi.org/10.1111/cogs.12235

Beck, S. R., & Riggs, K. J. (2014). Developing Thoughts About What Might Have Been. *Child Development Perspectives*, *8*(3), 175–179. https://doi.org/10.1111/cdep.12082

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's Thinking About Counterfactuals and Future Hypotheticals as Possibilities. *Child Development*, *77*(2), 413–426. https://doi.org/10.1111/j.1467-8624.2006.00879.x

Beran, M. J., & Hopkins, W. D. (2018). Self-Control in Chimpanzees Relates to General Intelligence. *Current Biology*, *28*(4), 574-579.e3. https://doi.org/10.1016/j.cub.2017.12.043

Bernstein, P. L. (1996). *Against the gods: The remarkable story of risk* (Vol. 383). John Wiley & Sons New York.

Bräuer, J., Kaminski, J., Riedel, J., Call, J., & Tomasello, M. (2006). Making inferences about the location of hidden food: Social dog, causal ape. *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *120*(1), 38–47. https://doi.org/10.1037/0735-7036.120.1.38

Broihanne, M.-H., Romain, A., Call, J., Thierry, B., Wascher, C. A. F., De Marco, A., Verrier, D., & Dufour, V. (2019). Monkeys (Sapajus apella and Macaca tonkeana) and great apes (Gorilla gorilla, Pongo abelii, Pan paniscus, and Pan troglodytes) play for the highest bid. *Journal of Comparative Psychology*, *133*(3), 301–312. https://doi.org/10.1037/com0000153

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126. https://doi.org/10.1016/j.neuron.2009.06.009

Brun, W., & Teigen, K. H. (1990). Prediction and postdiction preferences in guessing. *Journal of Behavioral Decision Making*, *3*(1), 17–28.

Byrne, R. M. J. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*, 135–157. https://doi.org/10.1146/annurev-psych-122414-033249

Cacchione, T., & Call, J. (2010). Intuitions about gravity and solidity in great apes: The tubes task. *Developmental Science*, *13*(2), 320–330. https://doi.org/10.1111/j.1467-7687.2009.00881.x

Call, J. (2004). Inferences about the location of food in the great apes (Pan paniscus, Pan troglodytes, Gorilla gorilla, and Pongo pygmaeus). *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *118*(2), 232–241. https://doi.org/10.1037/0735-7036.118.2.232

Call, J. (2006). Inferences by exclusion in the great apes: The effect of age and species. *Animal Cognition*, *9*(4), 393–403. https://doi.org/10.1007/s10071-006-0037-4

Call, J. (2010). Do apes know that they could be wrong? *Animal Cognition*, *13*(5), 689–700. https://doi.org/10.1007/s10071-010-0317-x

Call, J. (2022). The "avoid the empty cup" hypothesis does not explain great apes' (Gorilla gorilla, Pan paniscus, Pan troglodytes, Pongo abelii) responses in two three-cup one-item inference by exclusion tasks. *Journal of Comparative Psychology*, *136*, 172–188. https://doi.org/10.1037/com0000321

Call, J., & Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition*, *3*(4), 207–220. https://doi.org/10.1007/s100710100078

Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's Inference to the best explanation. *Synthese*, *180*(3), 419–442. https://doi.org/10.1007/s11229-009-9709-3

Carey, S., Leahy, B., Redshaw, J., & Suddendorf, T. (2020). Could It Be So? The Cognitive Science of Possibility. *Trends in Cognitive Sciences*, *24*(1), 3–4. https://doi.org/10.1016/j.tics.2019.11.007

Chua Chow, C., & Sarin, R. K. (2002). Known, Unknown, and Unknowable Uncertainties. *Theory and Decision*, *52*(2), 127–138. https://doi.org/10.1023/A:1015544715608

Comrie, A. E., Frank, L. M., & Kay, K. (2022). Imagination as a fundamental function of the hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1866), 20210336. https://doi.org/10.1098/rstb.2021.0336

Corballis, M. C. (2019). Mental time travel, language, and evolution. *Neuropsychologia*, *134*, 107202. https://doi.org/10.1016/j.neuropsychologia.2019.107202

Crimston, J., Redshaw, J., & Suddendorf, T. (2023). What are the odds? Preschoolers' ability to distinguish between possible, impossible, and probabilistically distinct future outcomes. *Developmental Psychology*, *59*(10), 1881–1891. https://doi.org/10.1037/dev0001587

Damerius, L. A., Graber, S. M., Willems, E. P., & van Schaik, C. P. (2017). Curiosity boosts orang-utan problem-solving ability. *Animal Behaviour*, *134*, 57–70. https://doi.org/10.1016/j.anbehav.2017.10.005

Danel, S., Chiffard-Carricaburu, J., Bonadonna, F., & Nesterova, A. P. (2021). Exclusion in the field: Wild brown skuas find hidden food in the absence of visual information. *Animal Cognition*, *24*(4), 867–876. https://doi.org/10.1007/s10071-021-01486-4

Danel, S., Rebout, N., Bonadonna, F., & Biro, D. (2022). Wild skuas can use acoustic cues to locate hidden food. *Animal Cognition*, *25*(5), 1357–1363. https://doi.org/10.1007/s10071-022-01611-x

Danel, S., Rebout, N., Osiurak, F., & Biro, D. (2022). Exclusion by donkey's ears: Donkeys (Equus asinus) use acoustic information to find hidden food in a two-way object-choice task. *Journal of Comparative Psychology*, *136*(1), 68–78. https://doi.org/10.1037/com0000308

De Petrillo, F., & Rosati, A. G. (2020). Logical inferences from visual and auditory information in ruffed lemurs and sifakas. *Animal Behaviour*, *164*, 193–204. https://doi.org/10.1016/j.anbehav.2020.03.010

De Waal, F. (2019, June 2). The Surprising Complexity of Animal Memories. *The Atlantic*. https://www.theatlantic.com/science/archive/2019/06/surprising-complexity-animal-memories/589420/

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, *13*(5), 798–803. https://doi.org/10.1111/j.1467-7687.2009.00943.x

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335–347. https://doi.org/10.1016/j.cognition.2013.12.001

Dequech, D. (2000). Fundamental Uncertainty and Ambiguity. *Eastern Economic Journal*, *26*(1), 41–60.

Diamond, A. (1990). Developmental Time Course in Human Infants and Infant Monkeys, and the Neural Bases of, Inhibitory Control in Reachinga. *Annals of the New York Academy of Sciences*, *608*(1), 637–676. https://doi.org/10.1111/j.1749-6632.1990.tb48913.x

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Douven, I. (2021). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/abduction/

Dubourg, E., & Baumard, N. (2022). Why imaginary worlds? The psychological foundations and cultural evolution of fictions with imaginary worlds. *Behavioral and Brain Sciences*, *45*, e276. https://doi.org/10.1017/S0140525X21000923

Duffrene, J., Petit, O., Thierry, B., Nowak, R., & Dufour, V. (2022). Both sheep and goats can solve inferential by exclusion tasks. *Animal Cognition*, *25*(6), 1631–1644. https://doi.org/10.1007/s10071-022-01656-y

Eckert, J., Call, J., Hermes, J., Herrmann, E., & Rakoczy, H. (2018). Intuitive statistical inferences in chimpanzees and humans follow Weber's law. *Cognition*, *180*, 99–107. https://doi.org/10.1016/j.cognition.2018.07.004

Eckert, J., Rakoczy, H., Call, J., Herrmann, E., & Hanus, D. (2018). Chimpanzees Consider Humans' Psychological States when Drawing Statistical Inferences. *Current Biology*, *28*(12), 1959-1963.e3. https://doi.org/10.1016/j.cub.2018.04.077

Edgington, D. (2011). 11 Causation First: Why Causation is Prior to Counterfactuals. In C. Hoerl, T. McCormack, & S. R. Beck (Eds.), *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199590698.003.0012

Eliaz, K., & Schotter, A. (2007). Experimental Testing of Intrinsic Preferences for NonInstrumental Information. *American Economic Review*, *97*(2), 166–169. https://doi.org/10.1257/aer.97.2.166

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms*. *The Quarterly Journal of Economics*, *75*(4), 643–669. https://doi.org/10.2307/1884324

Engelmann, J. M., Haux, L. M., Völter, C., Schleihauf, H., Call, J., Rakoczy, H., & Herrmann, E. (2023). Do chimpanzees reason logically? *Child Development*, *94*(5), 1102–1116. https://doi.org/10.1111/cdev.13861

Engelmann, J. M., Völter, C. J., Goddu, M. K., Call, J., Rakoczy, H., & Herrmann, E. (2023). Chimpanzees prepare for alternative possible outcomes. *Biology Letters*, *19*(6), 20230179. https://doi.org/10.1098/rsbl.2023.0179

Engelmann, J. M., Völter, C. J., O'Madagain, C., Proft, M., Haun, D. B. M., Rakoczy, H., & Herrmann, E. (2021). Chimpanzees consider alternative possibilities. *Current Biology*, *31*(20), R1377–R1378. https://doi.org/10.1016/j.cub.2021.09.012

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, *12*(2), 168–192. https://doi.org/10.1177/1088868308316091

Erdőhegyi, Á., Topál, J., Virányi, Z., & Miklósi, Á. (2007). Dog-logic: Inferential reasoning in a two-way choice task and its restricted use. *Animal Behaviour*, *74*(4), 725–737. https://doi.org/10.1016/j.anbehav.2007.03.004

Evans, J. S. B. (2006). *Dual system theories of cognition: Some issues*. *28*(28).

Evans, J. S. B. T. (2021). *The Rationality Debate in the Psychology of Reasoning: A Historical Review*. https://doi.org/10.7551/mitpress/11252.003.0008

Ferrigno, S., Huang, Y., & Cantlon, J. F. (2021). Reasoning Through the Disjunctive Syllogism in Monkeys. *Psychological Science*, *32*(2), 292–300. https://doi.org/10.1177/0956797620971653

Fichtel, C., Dinter, K., & Kappeler, P. M. (2020). The lemur baseline: How lemurs compare to monkeys and apes in the Primate Cognition Test Battery. *PeerJ*, *8*, e10025. https://doi.org/10.7717/peerj.10025

FitzGibbon, L., Komiya, A., & Murayama, K. (2021). The Lure of Counterfactual Curiosity: People Incur a Cost to Experience Regret. *Psychological Science*, *32*(2), 241–255. https://doi.org/10.1177/0956797620963615

FitzGibbon, L., Moll, H., Carboni, J., Lee, R., & Dehghani, M. (2019). Counterfactual curiosity in preschool children. *Journal of Experimental Child Psychology*, *183*, 146–157. https://doi.org/10.1016/j.jecp.2018.11.022

Fitzgibbon, L., & Murayama, K. (2022). Counterfactual curiosity: Motivated thinking about what might have been. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1866), 20210340. https://doi.org/10.1098/rstb.2021.0340

Floridi, L. (1997). *Scepticism and Animal Rationality: The Fortune of Chrysippus' Dog in the History of Western Thought*. *79*(1), 27–57. https://doi.org/10.1515/agph.1997.79.1.27

Fox, C. R., Goedde-Menke, M., & Tannenbaum, D. (2021). Ambiguity aversion and epistemic uncertainty. *Available at SSRN 3922716*.

Fox, C. R., & Tversky, A. (1995). Ambiguity Aversion and Comparative Ignorance. *The Quarterly Journal of Economics*, *110*(3), 585–603. https://doi.org/10.2307/2946693

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), Article 2. https://doi.org/10.1038/nrn2787

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement Learning or Active Inference? *PLoS ONE*, *4*(7), e6421. https://doi.org/10.1371/journal.pone.0006421

Gautam, S., Suddendorf, T., & Redshaw, J. (2021a). Do Monkeys and Young Children Understand Exclusive "Or" Relations? A Commentary on Ferrigno et al. (2021). *Psychological Science*, *32*(11), 1865–1867. https://doi.org/10.1177/09567976211024641

Gautam, S., Suddendorf, T., & Redshaw, J. (2021b). When can young children reason about an exclusive disjunction? A follow up to Mody and Carey (2016). *Cognition*, *207*, 104507. https://doi.org/10.1016/j.cognition.2020.104507

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

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), Article 12. https://doi.org/10.1038/s41583-018-0078-0

Grigoroglou, M., & Ganea, P. A. (2022). Language as a mechanism for reasoning about possibilities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1866), 20210334. https://doi.org/10.1098/rstb.2021.0334

Guerini, R., FitzGibbon, L., & Coricelli, G. (2020). The role of agency in regret and relief in 3- to 10-year-old children. *Journal of Economic Behavior & Organization*, *179*, 797–806. https://doi.org/10.1016/j.jebo.2020.03.029

Hanus, D., & Call, J. (2014). When maths trumps logic: Probabilistic judgements in chimpanzees. *Biology Letters*, *10*(12), 20140892. https://doi.org/10.1098/rsbl.2014.0892

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151. https://doi.org/10.1006/anbe.2000.1518

Harris, A. J. L., Rowley, M. G., Beck, S. R., Robinson, E. J., & McColgan, K. L. (2011). Agency Affects Adults', but not Children's, Guessing Preferences in a Game of Chance. *Quarterly Journal of Experimental Psychology*, *64*(9), 1772–1787. https://doi.org/10.1080/17470218.2011.582126

Haun, D. B. M., Nawroth, C., & Call, J. (2011). Great Apes' Risk-Taking Strategies in a Decision Making Task. *PLOS ONE*, *6*(12), e28801. https://doi.org/10.1371/journal.pone.0028801

Haux, L. M., Engelmann, J. M., Arslan, R. C., Hertwig, R., & Herrmann, E. (2023). Chimpanzee and Human Risk Preferences Show Key Similarities. *Psychological Science*, *34*(3), 358–369. https://doi.org/10.1177/09567976221140326

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, *4*(1), 5–28. https://doi.org/10.1007/BF00057884

Heilbronner, S. R., Rosati, A. G., Stevens, J. R., Hare, B., & Hauser, M. D. (2008). A fruit in the hand or two in the bush? Divergent risk preferences in chimpanzees and bonobos. *Biology Letters*, *4*(3), 246–249. https://doi.org/10.1098/rsbl.2008.0081

Heimbauer, L. A., Johns, T. N., & Weiss, D. J. (2019). Inferential reasoning in the visual and auditory modalities by cotton-top tamarins (Saguinus oedipus). *Journal of Comparative Psychology*, *133*(4), 496–501. https://doi.org/10.1037/com0000184

Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 805–812. https://doi.org/10.1037/a0018784

Herrmann, E., & Call, J. (2012). Are there geniuses among the apes? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2753–2761. https://doi.org/10.1098/rstb.2012.0191

Herrmann, E., Call, J., Hernàndez-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. https://doi.org/10.1126/science.1146282

Hill, A., Collier-Baker, E., & Suddendorf, T. (2011). Inferential reasoning by exclusion in great apes, lesser apes, and spider monkeys. *Journal of Comparative Psychology*, *125*(1), 91–103. https://doi.org/10.1037/a0020867

Hoerl, C., & McCormack, T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. *Behavioral and Brain Sciences*, *42*, e244. https://doi.org/10.1017/S0140525X18002157

Hopkins, W. D., Mareno, M. C., Webb, S. J. N., Schapiro, S. J., Raghanti, M. A., & Sherwood, C. C. (2021). Age-Related Changes in Chimpanzee (Pan troglodytes) Cognition: Cross-Sectional and Longitudinal Analyses. *American Journal of Primatology*, *83*(3), e23214. https://doi.org/10.1002/ajp.23214

Howell, W. C., & Burnett, S. A. (1978). Uncertainty measurement: A cognitive taxonomy. *Organizational Behavior and Human Performance*, *22*(1), 45–68. https://doi.org/10.1016/0030-5073(78)90004-1

Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J., & Dayan, P. (2016). The modulation of savouring by prediction error and its effects on choice. *eLife*, *5*, e13747. https://doi.org/10.7554/eLife.13747

Jelbert, S. A., Taylor, A. H., & Gray, R. D. (2015). Reasoning by exclusion in New Caledonian crows (Corvus moneduloides) cannot be explained by avoidance of empty containers. *Journal of Comparative Psychology*, *129*(3), 283–290. https://doi.org/10.1037/a0039313

Johnson, S. G., Merchant, T., & Keil, F. (2015). *Argument Scope in Inductive Reasoning: Evidence for an Abductive Account of Induction.* CogSci.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(43), 18243–18250. https://doi.org/10.1073/pnas.1012933107

Johnson-Laird, P. N. (2021). *Mental Models, Reasoning, and Rationality*. https://doi.org/10.7551/mitpress/11252.003.0014

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143–157. https://doi.org/10.1016/0010-0277(82)90023-3

Keupp, S., Grueneisen, S., Ludvig, E. A., Warneken, F., & Melis, A. P. (2021). Reduced risk-seeking in chimpanzees in a zero-outcome game. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1819), 20190673. https://doi.org/10.1098/rstb.2019.0673

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Knight, F. H. (1921). *Risk, uncertainty and profit* (Vol. 31). Houghton Mifflin.

Kruger, J., & Savitsky, K. (2004). The ?Reign of Error? In Social Psychology: On the Real Versus Imagined Consequences of Problem-Focused Research. *Behavioral and Brain Sciences*, *27*(3), 349–350. https://doi.org/10.1017/s0140525x04440084

Lacreuse, A., Parr, L., Chennareddi, L., & Herndon, J. G. (2018). Age-related decline in cognitive flexibility in female chimpanzees. *Neurobiology of Aging*, *72*, 83–88. https://doi.org/10.1016/j.neurobiolaging.2018.08.018

Lambert, M. L., & Osvath, M. (2018). Comparing chimpanzees' preparatory responses to known and unknown future outcomes. *Biology Letters*, *14*(9), 20180499. https://doi.org/10.1098/rsbl.2018.0499

Lampe, M., Bräuer, J., Kaminski, J., & Virányi, Z. (2017). The effects of domestication and ontogeny on cognition in dogs and wolves. *Scientific Reports*, *7*(1), 11690. https://doi.org/10.1038/s41598-017-12055-6

Leahy, B., Huemer, M., Steele, M., Alderete, S., & Carey, S. (2022). Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(52). https://doi.org/10.1073/pnas.2207499119

Leahy, B. P., & Carey, S. E. (2020). The Acquisition of Modal Concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78. https://doi.org/10.1016/j.tics.2019.11.004

Leahy, B., Rafetseder, E., & Perner, J. (2014). Basic Conditional Reasoning: How Children Mimic Counterfactual Reasoning. *Studia Logica*, *102*(4), 793–810. https://doi.org/10.1007/s11225-013-9510-7

Lipton, P. (2000). Inference to the Best Explanation. In W. Newton-Smith (Ed.), *A Companion to the Philosophy of Science* (Vol. 18). Blackwell Oxford.

Lipton, P. (2017). Inference to the best explanation. *A Companion to the Philosophy of Science*, 184–193.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75. https://doi.org/10.1037/0033-2909.116.1.75

Mahr, J. B. (2020). The dimensions of episodic simulation. *Cognition*, *196*, 104085. https://doi.org/10.1016/j.cognition.2019.104085

Maille, A., & Roeder, J. J. (2012). Inferences about the location of food in lemurs (Eulemur macaco and Eulemur fulvus): A comparison with apes and monkeys. *Animal Cognition*, *15*(6), 1075–1083. https://doi.org/10.1007/s10071-012-0531-9

Marín Manrique, H., & Call, J. (2015). Age-dependent cognitive inflexibility in great apes. *Animal Behaviour*, *102*, 1–6. https://doi.org/10.1016/j.anbehav.2015.01.002

Marsh, H. L., Vining, A. Q., Levendoski, E. K., & Judge, P. G. (2015). Inference by exclusion in lion-tailed macaques (Macaca silenus), a hamadryas baboon (Papio hamadryas), capuchins (Sapajus apella), and squirrel monkeys (Saimiri sciureus). *Journal of Comparative Psychology*, *129*(3), 256–267. https://doi.org/10.1037/a0039316

Mcauliffe, W. H. B. (2015). How did Abduction Get Confused with Inference to the Best Explanation? *Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy*, *51*(3), 300–319.

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The development of counterfactual reasoning about doubly-determined events. *Cognitive Development*, *45*, 1–9. https://doi.org/10.1016/j.cogdev.2017.10.001

McCormack, T., & Hoerl, C. (2017). The Development of Temporal Concepts: Learning to Locate Events in Time. *Timing & Time Perception*, *5*(3–4), 297–327. https://doi.org/10.1163/22134468-00002094

Melis, A. P., Hare, B., & Tomasello, M. (2006). Chimpanzees Recruit the Best Collaborators. *Science*, *311*(5765), 1297–1300. https://doi.org/10.1126/science.1123007

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Mikolasch, S., Kotrschal, K., & Schloegl, C. (2012). Is caching the key to exclusion in corvids? The case of carrion crows (Corvus corone corone). *Animal Cognition*, *15*(1), 73–82. https://doi.org/10.1007/s10071-011-0434-1

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40–48. https://doi.org/10.1016/j.cognition.2016.05.012

Mulcahy, N. J., & Call, J. (2006). How great apes perform on a modified trap-tube task. *Animal Cognition*, *9*(3), 193–199. https://doi.org/10.1007/s10071-006-0019-6

Nawroth, C., Borell, E. von, & Langbein, J. (2014). Exclusion Performance in Dwarf Goats (Capra aegagrus hircus) and Sheep (Ovis orientalis aries). *PLOS ONE*, *9*(4), e93534. https://doi.org/10.1371/journal.pone.0093534

Nawroth, C., & von Borell, E. (2015). Domestic pigs' (Sus scrofa domestica) use of direct and indirect visual and auditory cues in an object choice task. *Animal Cognition*, *18*(3), 757–766. https://doi.org/10.1007/s10071-015-0842-8

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4- and 5-year-olds. *Cognition*, *183*, 57–66. https://doi.org/10.1016/j.cognition.2018.10.027

Nyhout, A., Henke, L., & Ganea, P. A. (2019). Children's Counterfactual Reasoning About Causally Overdetermined Events. *Child Development*, *90*(2), 610–622. https://doi.org/10.1111/cdev.12913

O'Hara, M., Auersperg, A. M. I., Bugnyar, T., & Huber, L. (2015). Inference by Exclusion in Goffin Cockatoos (Cacatua goffini). *PLOS ONE*, *10*(8), e0134894. https://doi.org/10.1371/journal.pone.0134894

O'Hara, M., Schwing, R., Federspiel, I., Gajdon, G. K., & Huber, L. (2016). Reasoning by exclusion in the kea (Nestor notabilis). *Animal Cognition*, *19*, 965–975. https://doi.org/10.1007/s10071-016-0998-x

O'Madagain, C., Helming, K. A., Schmidt, M. F. H., Shupe, E., Call, J., & Tomasello, M. (2022). Great apes and human children rationally monitor their decisions. *Proceedings*

of the Royal Society B: Biological Sciences, *289*(1971), 20212686. https://doi.org/10.1098/rspb.2021.2686

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

Paukner, A., Anderson, J. R., & Fujita, K. (2006). Redundant food searches by capuchin monkeys (Cebus apella): A failure of metacognition? *Animal Cognition*, *9*(2), 110–117. https://doi.org/10.1007/s10071-005-0007-2

Paukner, A., Huntsberry, M. E., & Suomi, S. J. (2009). Tufted capuchin monkeys (Cebus apella) spontaneously use visual but not acoustic information to find hidden food items. *Journal of Comparative Psychology*, *123*(1), 26–33. https://doi.org/10.1037/a0013128

Pepperberg, I. M., Koepke, A., Livingston, P., Girard, M., & Hartsfield, L. A. (2013). Reasoning by inference: Further studies on exclusion in grey parrots (Psittacus erithacus). *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *127*(3), 272–281. https://doi.org/10.1037/a0031641

Perner, J., Sprung, M., & Steinkogler, B. (2004). Counterfactual conditionals and false belief: A developmental dissociation. *Cognitive Development*, *19*(2), 179–201. https://doi.org/10.1016/j.cogdev.2003.12.001

Petit, O., Dufour, V., Herrenschmidt, M., De Marco, A., Sterck, E. H. M., & Call, J. (2015). Inferences about food location in three cercopithecine species: An insight into the socioecological cognition of primates. *Animal Cognition*, *18*(4), 821–830. https://doi.org/10.1007/s10071-015-0848-2

Plotnik, J. M., Pokorny, J. J., Keratimanochaya, T., Webb, C., Beronja, H. F., Hennessy, A., Hill, J., Hill, V. J., Kiss, R., Maguire, C., Melville, B. L., Morrison, V. M. B., Seecoomar, D., Singer, B., Ukehaxhaj, J., Vlahakis, S. K., Ylli, D., Clayton, N. S., Roberts, J., … Getz, D. (2013). Visual Cues Given by Humans Are Not Sufficient for

Asian Elephants (Elephas maximus) to Find Hidden Food. *PLOS ONE*, *8*(4), e61174. https://doi.org/10.1371/journal.pone.0061174

Plotnik, J. M., Shaw, R. C., Brubaker, D. L., Tiller, L. N., & Clayton, N. S. (2014). Thinking with their trunks: Elephants use smell but not sound to locate food and exclude nonrewarding alternatives. *Animal Behaviour*, *88*, 91–98. https://doi.org/10.1016/j.anbehav.2013.11.011

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual Reasoning: Developing a Sense of "Nearest Possible World". *Child Development*, *81*(1), 376–389. https://doi.org/10.1111/j.1467-8624.2009.01401.x

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389–404. https://doi.org/10.1016/j.jecp.2012.10.010

Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition*, *131*(1), 60–68. https://doi.org/10.1016/j.cognition.2013.12.011

Rathke, E.-M., & Fischer, J. (2020). Differential ageing trajectories in motivation, inhibitory control and cognitive flexibility in Barbary macaques (Macaca sylvanus). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1811), 20190617. https://doi.org/10.1098/rstb.2019.0617

Read, D. W., Manrique, H. M., & Walker, M. J. (2022). On the working memory of humans and great apes: Strikingly similar or remarkably different? *Neuroscience & Biobehavioral Reviews*, *134*, 104496. https://doi.org/10.1016/j.neubiorev.2021.12.019

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, *17*(3), Article 3. https://doi.org/10.1038/nrn.2015.30

Redshaw, J., Leamy, T., Pincus, P., & Suddendorf, T. (2018). Young children's capacity to imagine and prepare for certain and uncertain future outcomes. *PloS One*, *13*(9), e0202606. https://doi.org/10.1371/journal.pone.0202606

Redshaw, J., & Suddendorf, T. (2016). Children's and Apes' Preparatory Responses to Two Mutually Exclusive Possibilities. *Current Biology*, *26*(13), 1758–1762. https://doi.org/10.1016/j.cub.2016.04.062

Redshaw, J., & Suddendorf, T. (2020). Temporal Junctures in the Mind. *Trends in Cognitive Sciences*, *24*(1), 52–64. https://doi.org/10.1016/j.tics.2019.10.009

Redshaw, J., Suddendorf, T., Neldner, K., Wilks, M., Tomaselli, K., Mushin, I., & Nielsen, M. (2019). Young Children From Three Diverse Cultures Spontaneously and Consistently Prepare for Alternative Future Possibilities. *Child Development*, *90*(1), 51–61. https://doi.org/10.1111/cdev.13084

Rescorla, M. (2009). Chrysippus' dog as a case study in non-linguistic cognition. *The Philosophy of Animal Minds*, 52–71.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In *Simple heuristics that make us smart* (pp. 141–167). Oxford University Press.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, *12*(2), 129–134. https://doi.org/10.1111/1467-9280.00322

Rivas-Blanco, D., Krause, S. D., Marshall-Pescini, S., & Range, F. (2024). *Inference in wolves and dogs: The "cups task", revisited* (p. 2024.09.03.610928). bioRxiv. https://doi.org/10.1101/2024.09.03.610928

Rivière, J., Kurt, A., & Meunier, H. (2019). Choice under risk of gain in tufted capuchin monkeys (Sapajus apella): A comparison with young children (Homo sapiens) and mangabey monkeys (Cercocebus torquatus torquatus). *Journal of Neuroscience, Psychology, and Economics*, *12*(3–4), 159.

Rivière, J., Stomp, M., Augustin, E., Lemasson, A., & Blois-Heulin, C. (2018). Decision-making under risk of gain in young children and mangabey monkeys. *Developmental Psychobiology*, *60*(2), 176–186. https://doi.org/10.1002/dev.21592

Robinson, E. J., & Beck, S. (2000). What is difficult about counterfactual reasoning? In *Children's reasoning and the mind* (pp. 101–119). Psychology Press/Taylor & Francis (UK).

Robinson, E. J., Pendle, J. E. C., Rowley, M. G., Beck, S. R., & McColgan, K. L. T. (2009). Guessing imagined and live chance events: Adults behave like children with live events. *British Journal of Psychology*, *100*(4), 645–659. https://doi.org/10.1348/000712608X386810

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's Sensitivity to Their Own Relative Ignorance: Handling of Possibilities Under Epistemic and Physical Uncertainty. *Child Development*, *77*(6), 1642–1655. https://doi.org/10.1111/j.1467-8624.2006.00964.x

Roig, A., Meunier, H., Poulingue, E., Marty, A., Thouvarecq, R., & Rivière, J. (2022). Is economic risk proneness in young children (Homo sapiens) driven by exploratory behavior? A comparison with capuchin monkeys (Sapajus apella). *Journal of Comparative Psychology*, *136*(2), 140–150. https://doi.org/10.1037/com0000314

Rosati, A. G., & Hare, B. (2010). Chimpanzees and bonobos distinguish between risk and ambiguity. *Biology Letters*, *7*(1), 15–18. https://doi.org/10.1098/rsbl.2010.0927

Rosati, A. G., & Hare, B. (2012). Decision making across social contexts: Competition increases preferences for risk in chimpanzees and bonobos. *Animal Behaviour*, *84*(4), 869–879. https://doi.org/10.1016/j.anbehav.2012.07.010

Rosati, A. G., & Hare, B. (2013). Chimpanzees and Bonobos Exhibit Emotional Responses to Decision Outcomes. *PLOS ONE*, *8*(5), e63058. https://doi.org/10.1371/journal.pone.0063058

Rosenberg, R. D., & Feigenson, L. (2013). Infants hierarchically organize memory representations. *Developmental Science*, *16*(4), 610–621. https://doi.org/10.1111/desc.12055

Sabbatini, G., & Visalberghi, E. (2008). Inferences about the location of food in capuchin monkeys (Cebus apella) in two sensory modalities. *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *122*(2), 156–166. https://doi.org/10.1037/0735-7036.122.2.156

Schloegl, C., Dierks, A., Gajdon, G. K., Huber, L., Kotrschal, K., & Bugnyar, T. (2009). What You See Is What You Get? Exclusion Performances in Ravens and Keas. *PLOS ONE*, *4*(8), e6368. https://doi.org/10.1371/journal.pone.0006368

Schloegl, C., Schmidt, J., Boeckle, M., Weiß, B. M., & Kotrschal, K. (2012). Grey parrots use inferential reasoning based on acoustic cues alone. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1745), 4135–4142. https://doi.org/10.1098/rspb.2012.1292

Schmitt, V., & Fischer, J. (2009). Inferential reasoning and modality dependent discrimination learning in olive baboons (Papio hamadryas anubis). *Journal of Comparative Psychology*, *123*(3), 316–325. https://doi.org/10.1037/a0016218

Schurz, G. (2021). *Evolution of Rationality*. https://doi.org/10.7551/mitpress/11252.003.0009

Seed, A. M., Call, J., Emery, N. J., & Clayton, N. S. (2009a). Chimpanzees solve the trap problem when the confound of tool-use is removed. *Journal of Experimental Psychology. Animal Behavior Processes*, *35*(1), 23–34. https://doi.org/10.1037/a0012925

Seed, A. M., Call, J., Emery, N. J., & Clayton, N. S. (2009b). Chimpanzees solve the trap problem when the confound of tool-use is removed. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(1), 23–34. https://doi.org/10.1037/a0012925

Seed, A. M., Tebbich, S., Emery, N. J., & Clayton, N. S. (2006). Investigating Physical Cognition in Rooks, Corvus frugilegus. *Current Biology*, *16*(7), 697–701. https://doi.org/10.1016/j.cub.2006.02.066

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating Into the Future or Driven by the Past. *Perspectives on Psychological Science*, *8*(2), 119–141. https://doi.org/10.1177/1745691612474317

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning.* Lawrence Erlbaum Associates Publishers.

Stigler, G. J. (1961). The Economics of Information. *Journal of Political Economy*, *69*(3), 213–225. https://doi.org/10.1086/258464

Suddendorf, T., & Busby, J. (2005). Making decisions with the future in mind: Developmental and comparative identification of mental time travel. *Learning and Motivation*, *36*(2), 110–125. https://doi.org/10.1016/j.lmot.2005.02.010

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *The Behavioral and Brain Sciences*, *30*(3), 299–313; discussion 313-351. https://doi.org/10.1017/S0140525X07001975

Suddendorf, T., Crimston, J., & Redshaw, J. (2017). Preparatory responses to socially determined, mutually exclusive possibilities in chimpanzees and children. *Biology Letters*, *13*(6), 20170170. https://doi.org/10.1098/rsbl.2017.0170

Suddendorf, T., Watson, K., Bogaart, M., & Redshaw, J. (2020). Preparation for certain and uncertain future outcomes in young children and three species of monkey. *Developmental Psychobiology*, *62*(2), 191–201. https://doi.org/10.1002/dev.21898

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). Oxford University Press.

Ülkümen, G., Fox, C. R., & Malle, B. F. (2016). Two dimensions of subjective uncertainty: Clues from natural language. *Journal of Experimental Psychology: General*, *145*(10), 1280–1297. https://doi.org/10.1037/xge0000202

Vasconcelos, M., Monteiro, T., & Kacelnik, A. (2015). Irrational choice and the value of information. *Scientific Reports*, *5*, 13874. https://doi.org/10.1038/srep13874

Völter, C. J., & Call, J. (2014a). 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. https://doi.org/10.1037/a0035434

Völter, C. J., & Call, J. (2014b). The cognitive underpinnings of flexible tool use in great apes. *Journal of Experimental Psychology. Animal Learning and Cognition*, *40*(3), 287–302. https://doi.org/10.1037/xan0000025

Völter, C. J., & Call, J. (2017). Causal and inferential reasoning in animals. In J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, & T. Zentall (Eds.), *APA handbook of comparative psychology: Perception, learning, and cognition.* (pp. 643–671). American Psychological Association. https://doi.org/10.1037/0000012-029

Völter, C. J., Mundry, R., Call, J., & Seed, A. M. (2019). Chimpanzees flexibly update working memory contents and show susceptibility to distraction in the self-ordered search task. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907), 20190715. https://doi.org/10.1098/rspb.2019.0715

Völter, C. J., Reindl, E., Felsche, E., Civelek, Z., Whalen, A., Lugosi, Z., Duncan, L., Herrmann, E., Call, J., & Seed, A. M. (2022). The structure of executive functions in preschool children and chimpanzees. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-08406-7

Völter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2018). Comparative psychometrics: Establishing what differs is central to understanding what evolves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170283. https://doi.org/10.1098/rstb.2017.0283

Wang, M. Z., & Hayden, B. Y. (2019). Monkeys are curious about counterfactual outcomes. *Cognition*, *189*, 1–10. https://doi.org/10.1016/j.cognition.2019.03.009

Watson, J. S., Gergely, G., Csanyi, V., Topal, J., Gacsi, 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. *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *115*(3), 219–226. https://doi.org/10.1037/0735-7036.115.3.219

Weisberg, D. S., & Gopnik, A. (2013). Pretense, Counterfactuals, and Bayesian Causal Models: Why What Is Not Real Really Matters. *Cognitive Science*, *37*(7), 1368–1381. https://doi.org/10.1111/cogs.12069

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015. https://doi.org/10.1073/pnas.0704450105

Appendix I: Supplementary data.

8.1 Supplementary data to chapter 2

Table 8.1.1: Effect sizes from the model to predict taking the half grape in experiment 2.

	χ^2	df	p-value
Condition	79.421	1	<.001
Age^2	10.995	2	0.004
Species	3.947	3	0.267
Block	5.659	1	0.017
Condition: Age^2	13.145	2	0.001
Certain: Species	8.314	3	0.040
Certain: Block	0.196	1	0.658

.

Table 8.1.2: Coefficients from the final model to predict taking the half grape in experiment 2 with the reference level set to occluded

	β	$CI_{2.5}$	$CI_{97.5}$	p-value
(Intercept)	-0.382	-0.932	0.168	0.173
Visible	-2.134	-2.838	-1.43	<.001
Age	-2.061	-13.888	9.766	0.733
Age^2	20.96	10.277	31.642	<.001
Chimpanzee	-0.373	-1.314	0.568	0.438
Gorilla	-0.448	-1.348	0.452	0.329
Orangutan	0.047	-0.903	0.996	0.923
Block	-0.458	-0.858	-0.057	0.025
Visible: Age	19.178	5.933	32.423	0.005
Visible: Age^2	-12.1	-22.957	-1.244	0.029
Visible: Chimpanzee	0.064	-1.079	1.207	0.913
Visible: Gorilla	0.06	-1.091	1.211	0.919
Visible: Orangutan	1.287	0.32	2.254	0.009
Visible: block	0.163	-0.558	0.884	0.658

Table 8.1.3: Coefficients from refitting the model in experiment 2 without Coco

	β	$CI_{2.5}$	$CI_{97.5}$	p-value
(Intercept)	-2.276	-3.341	-1.210	0.000
Visible	2.236	1.063	3.409	< .001
Age	13.511	0.089	26.932	0.048
Age^2	12.391	-0.629	25.412	0.062
Chimpanzee	-0.167	-1.387	1.052	0.788
Gorilla	-0.505	-1.763	0.754	0.432
Orangutan	1.346	0.273	2.418	0.014
Block	-0.321	-0.956	0.313	0.321
Occluded: Age	-21.249	-33.632	-8.866	0.001
Occluded: Age^2	6.817	-5.289	18.923	0.270
Occluded: Chimpanzee	-0.081	-1.225	1.063	0.890
Occluded: Gorilla	-0.007	-1.204	1.190	0.991
Occluded: Orangutan	-1.291	-2.260	-0.322	0.009
Occluded: block	-0.110	-0.862	0.642	0.774

8.2 Supplementary data to Chapter 3

Table 8.2.1: Effect sizes from the model to predict taking the half-piece in experiment 2.

	χ^2	df	p-value
Trial Type	56.073	1	< .001
Age	1.560	1	0.212
Base rate	13.347	1	< .001
Trial Number	0.910	1	0.340
Species	4.228	3	0.238
Trial Type: Age	2.943	1	0.086
Trial Type: Base Rate	6.979	1	<.001
Trial Type: Trial Number	0.510	1	0.475
Trial Type: Species	7.807	3	0.050

*Table 8.2.2: Coefficients of the GLMM model fitted to predict taking the half grape in Experiment 3.(half ~ remove + remove*session + (1|id) , family = binomial()*

Term	Estimate	CI 2.5%	CI 97.5%	P-value
(Intercept)	-0.107	-1.424	-1.424	.874
Remove Empty	-3.327	-5.721	-5.721	.006
Session	0.69	-0.198	-0.198	.128
Remove Empty:Session	0.437	-1.062	-1.062	.568

8.3 Supplementary data to Chapter 4

Table 8.3.1: Pairwise comparisons for Figure 4.4.1. The top half of the table shows differences by trial type for the same condition, the bottom half shows differences by condition for the same trial type.

Experiment	Location	Factor	Group1	Group2	n1	n2	statistic	df	p	p.signif
One-choice	BRU	Control	Reveal Empty	Reveal Baited	6	6	5.394	5	0.003	**
		Standard	Reveal Empty	Reveal Baited	6	6	1.574	5	0.176	ns
	TWZ	Control	Reveal Empty	Reveal Baited	9	9	6.021	8	< 0.001	***
		Standard	Reveal Empty	Reveal Baited	9	9	0.577	8	0.580	ns
		E1	Reveal Empty	Reveal Baited	9	9	-2.055	8	0.074	ns
Two-choice	BRU	Control	Reveal Empty	Reveal Baited	5	6	0.158	4	0.882	ns
		Standard	Reveal Empty	Reveal Baited	6	5	2.201	4	0.093	ns
	TWZ	E1	Reveal Empty	Reveal Baited	9	9	-3.499	8	0.008	**
		Control	Reveal Empty	Reveal Baited	7	9	3.509	6	0.013	*
		Standard	Reveal Empty	Reveal Baited	9	9	-0.215	8	0.835	ns
One-choice	BRU	Reveal Empty	Control	Standard	6	6	1.348	5	0.235	ns
		Reveal Baited	Control	Standard	6	6	1.038	5	0.347	ns
	TWZ	Reveal Empty	Experiment 1	Standard	9	9	0.590	8	0.572	ns
		Reveal Baited	Experiment 1	Standard	9	9	1.974	8	0.084	ns
		Reveal Empty	Control	Standard	9	9	4.990	8	0.001	**
		Reveal Baited	Control	Standard	9	9	-0.668	8	0.523	ns
Two-choice	BRU	Reveal Empty	Control	Standard	5	6	0.432	4	0.688	ns
		Reveal Baited	Control	Standard	6	5	4.081	4	0.015	*
	TWZ	Reveal Empty	Experiment 1	Standard	9	9	-2.627	8	0.030	*
		Reveal Baited	Experiment 1	Standard	9	9	-0.721	8	0.492	ns
		Reveal Empty	Control	Standard	7	9	1.106	6	0.311	ns
		Reveal Baited	Control	Standard	9	9	-1.682	8	0.131	ns

Table 8.3.2: Error rates for two-choice trials of experiment 2.

Trial Type	Location	Species	ID	Correct Both	Incorrect Both	Incorrect First	Incorrect Second
Standard	BRU	Chimpanzee	Edith	0	0	1	0
			Eva	0	0.333	0.5	0.167
			Frek	0	0	0.667	0.333
			Kilimi	0.167	0	0.5	0.333
			Qafzeh	0	0.333	0.167	0.5
			Velu	0.167	0.333	0.167	0.333
	TWZ	Bonobo	Likemba	0.167	0.167	0.333	0.333
			Lola	0.5	0.167	0.167	0.167
			Lucuma	0.333	0.167	0.5	0
			Malaika	0.2	0	0.2	0.6
			Ndeko	0.2	0.4	0.2	0.2
		Gorilla	Lope	0.167	0	0.5	0.333
			Shufai	0.667	0.167	0.167	0
		Orangutan	Batu	0.333	0.167	0.167	0.333
			Kayan	0.167	0.167	0.667	0
Control	BRU	Chimpanzee	Edith	0.25	0.25	0	0.5
			Eva	0.25	0.167	0.167	0.417
			Frek	0.5	0	0	0.5
			Kilimi	0.333	0	0.083	0.583
			Qafzeh	0.417	0.083	0.083	0.417
			Velu	0.5	0	0.167	0.333
	TWZ	Bonobo	Likemba	0.75	0	0.25	0
			Lola	0.5	0.083	0.333	0.083
			Lucuma	0.667	0	0.167	0.167
			Malaika	0.5	0.167	0.083	0.25
			Ndeko	0.333	0	0.083	0.583
		Gorilla	Lope	0.417	0	0	0.583
			Shufai	0.167	0.083	0.417	0.333
		Orangutan	Batu	0.25	0	0.25	0.5
			Kayan	0.5	0	0	0.5

8.4 Supplementary data to chapter 5

Table 8.4.1: Individual rates of checking each tube in experiment 2.

ID	Check Neither	Check Unchosen	Check Unavailable	Check Both	Unchosen First	Unavailable First
Eva	0	0.783	0.826	0.609	0.609	0.391
Frek	0	0.875	0.875	0.75	0.292	0.708
Kilimi	0	0.625	0.833	0.458	0.208	0.792
Masindi	0	0.875	0.792	0.667	0.458	0.542
Paul	0.75	0	0.25	0	0	0.25
Qafzeh	0	1	0.875	0.875	0.458	0.542
Rene	0	0.625	0.625	0.25	0.5	0.5
Velu	0	1	1	1	0.5	0.5

Appendix II: Ethical approval forms.